DOCUMENT RESUME

ED 165 481 FL 010 007

AUTHOR Spolsky, Bernard, Ed.
TITLE Approaches to Language Testing. Advances in Language Testing Series: 2. Papers in Applied Linguistics.
INSTITUTION Center for Applied Linguistics, Arlington, Va.
PUB DATE Aug 78
NOTE 81p.
AVAILABLE FROM Center for Applied Linguistics, 1611 N. Kent St., Arlington, Va. 22209 ($4.95)

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS Applied Linguistics; Diagnostic Tests; Educational Psychology; Educational Theories; *Language Proficiency; *Language Tests; Linguistic Theory; *Measurement Techniques; Pragmatics; Prognostic Tests; Psycholinguistics; *Psychometrics; Sociolinguistics; *Testing

ABSTRACT

This volume, one in a series on modern language testing, collects four essays dealing with current approaches to language testing. The introduction traces the development of language testing theory and examines the role of linguistics in this area. "The Psycholinguistic Basis," by E. Ingram, discusses some interpretations of the term "psycholinguistics" and relates them to traditional and recent language testing practices. "Psychometric Considerations in Language Testing," by J.L.D. Clark, discusses aspects of psychometric practice with regard to three broad categories of purpose within the area of language testing: (1) prognostic testing, (2) diagnostic testing, and (3) proficiency testing. "The Sociolinguistic Foundations of Language Testing," by J.A. Fishman and R.L. Cooper, illustrates the usefulness of a sociolinguistic approach and provides justification for it by the construction of language assessment procedures. "Pragmatics and Language Testing," by J.W. Oller, Jr., discusses in historical perspective the major concepts of pragmatics and relates them to language testing. (AM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED165481

Papers in Applied Linguistics

ADVANCES IN LANGUAGE TESTING SERIES: 2

Approaches to Language Testing
edited by Bernard Spolsky

Bernard Spolsky, General Series Editor

FL010007

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Center for Applied Linguistics

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Center for Applied Linguistics

The editor wishes to express his thanks to all the authors who have contributed to this volume. He also wishes to express his special appreciation and thanks to Marcia E. Taylor of the Center's Communication & Publications Program for her careful and painstaking help in the copy-editing of this volume and in its final preparation for publication.

August 1978
Copyright © 1978
by the Center for Applied Linguistics
1611 North Kent Street
Arlington, Virginia 22209

ISBN: 87281-07505

Library of Congress Catalog Card Number: 78-62080

Printed in the United States of America

Contents

Preface iv

Introduction: Linguists and Language Testers v
Bernard Spolsky, University of New Mexico

The Psycholinguistic Basis 1
Elisabeth Ingram, University of Edinburgh and University of Trondheim

Psychometric Considerations in Language Testing 15
John L.D. Clark, Educational Testing Service

The Sociolinguistic Foundations of Language Testing 31
Joshua A. Fishman, Yeshiva University, and Robert L. Cooper, Hebrew University

Pragmatics and Language Testing 39
John W. Oller, Jr., University of New Mexico

References 59

Preface

Language testing is one of the most obvious and important areas for activity in applied linguistics. It fits the general paradigm well: the problem is clearly language related, and the solution must come from linguistics and from another discipline--psychometrics--as well. Each of the major branches of linguistics--theoretical, psycholinguistic, and sociolinguistic--has its own special role to play, and each has exerted its influence on the development of the field. The involvement of psychometrics is a necessity as well. Good language testing needs to be based on relevant knowledge from applied linguistics and from psychometrics.

Given this, and considering the social relevance of the field, it is appropriate that this third series of *Papers in Applied Linguistics* be dedicated to chronicling *Advances in Language Testing*. The series will first survey the state of the art and then present theoretical, practical, and technical articles that record its progress. Each issue will have a specific theme; the series as a whole is meant to provide a means for continuing communication among all those concerned with language testing, whether as users, practitioners, or theorists.

The first three fasicles have a special history which should be mentioned here. They include original and revised versions of articles commissioned to appear in a volume intended to be called *Current Trends in Language Testing*. The original publisher's difficulties left the manuscripts in limbo for some time, and the size of the enterprise discouraged other publishers from taking up the project. Written as a survey of the state of the art, they are a good start for the present series.

Bernard Spolsky
April 1978

Introduction: Linguists and Language Testers

One of the distinguishing features of Western education in the twentieth century has been the emphasis on testing. Formal objective testing has come to be considered one of the most critical steps in the modernization of education. In many countries, the conflict between traditional, "subjective" examinations and the newer "objective" standardized tests is still a central issue for professional and public debate. Concurrently with the growth of testing, there has developed a body of professionals trained and qualified in educational measurement. These testers, whose field is called psychometrics, tend to find their basic concepts and techniques in psychology, in general, and in educational psychology, in particular. This perhaps explains their special concern with the question of *how* to test: treated often as technicians, they tend to assume less responsibility for deciding *what* to test. While this is true of testing in most fields, in the area of language testing the situation is fundamentally different: language testers are much more likely to be linguists, and thus subject-matter specialists, than they are to be trained primarily in educational measurement. The tendency is clearly illustrated in this volume, where most of the contributors, all of them professionally involved with language testing to a great extent, would consider themselves linguists rather than psychologists. It will be the purpose of this introduction to attempt to find some explanation for this trend, and to explore the special reasons why applied linguists find language testing so fruitful an activity for their research and practice.

It is useful, though an over-generalization, to divide language testing into three major trends, which I will label the pre-scientific, the psychometric-structuralist, and the integrative-sociolinguistic. The trends follow in order but overlap in time and approach. The third picks up many elements of the first, and the second and third co-exist and compete. But the crude classifications will provide a framework for discussion and some notion of progress in the field.

The pre-scientific period (or trend, for it still holds sway in many parts of the world) may be characterized by lack of concern for statistical matters or for such notions as objectivity and reliability. In its simplest form, it assumes that one can and must rely completely on the judgment of an experienced teacher, who can tell after a few minutes' conversation, or after reading a student's essay, what mark to give. In the pre-scientific mode, oral examinations of any kind were the exception: language testing was assumed to be a matter of open-ended written examinations. Depending on the language teaching philosophy, such examinations would consist of passages for translation into or from the foreign language; free composition in it; and selected items of grammatical, textual, or cultural interest. During this period, and in this approach, language tests are clearly the business of language teachers, or, in more formal situations, of language teachers promoted or specially appointed as examiners. No special expertise is required: if a person knows how to teach,

it is to be assumed that he can judge the proficiency of his students.

The next period, however, sees the invasion of the field by experts. The psychometric-structuralist trend, though hyphenated for reasons that will become apparent, is marked by the interaction (and conflict) of two sets of experts, agreeing with each other mainly in their belief that testing can be made precise, objective, reliable, and scientific. The first of these groups of experts were the testers, the psychologists responsible for the development of modern theories and techniques of educational measurement. Their key concerns have been to provide "objective" measures using various statistical techniques to assure reliability and certain kinds of validity. Their first thrust was to demonstrate the unreliability of traditional examinations, and studies such as those of Pilliner (1952) and others on the marking of essays showed how unreliable subjective scores can be. This done, they moved to develop more reliable measures, working to find either techniques for making judgments more reliable or new kinds of test items more amenable to control.

The better known work of the testers was the development of short item, multiple choice, "objective" tests. The demands of statistical measures of reliability and validity were seen as of paramount importance: "Firstly, the shape of all tests, whether predictive or non-predictive, language or non-language, is primarily determined by the need to test the tests for reliability and validity. That is why, for instance, the multiple choice technique of answering is so common (Ingram, 1968, p. 74)."

There were two results from this emphasis. First, tests like this required written response, and so were limited to reading and listening. Second, the items chosen did not reflect newer ideas about language teaching and learning. The testers and psychologists added "scientific" techniques to language testing, but left a great number of deficiencies. Writing in 1952, John Carroll quotes with approval the criticism of language testing that Robert Lado had just made in the summary of his doctoral thesis:

> A number of conclusions are reached. They are (1) that a great lag exists in measurement in English as a foreign language, (2) that the lag is connected with unscientific views of language, (3) that the science of language should be used in defining what to teach....The study gives procedures for the application of linguistics to the development of foreign languages tests (Carroll, 1953).

Carroll confirms Lado's judgments and adds "that a great lag exists in all foreign language measurement." The lag had shown up first in the study of foreign language teaching carried out by Agard and Dunkel (1948): the only tests available were written tests of vocabulary, reading, and grammar, and none were good at measuring the aural-oral skills that were coming to be emphasized.

The second major impetus of the "scientific" period, or approach, then, was when a new set of experts added notions from the science of language to those from the science of educational measurement. One scholar who has straddled the two fields for most of his career is John B. Carroll whose early (1940) and recent (1978) work alike show his concern with psychologically and linguistically valid measures of verbal abilities, whether in native or learned languages. Carroll's special role in the development of language tests has arisen from his ability to speak as a fellow professional to both linguists and psychologists, and his influence has been widely felt (cf. Carroll, 1968a, 1968b, 1972b). The importance of his activities will be clear after a brief glance at

lists of references in a volume such as this.

In reviews of the state of the art written in 1952 (Carroll, 1952, 1953), Carroll drew attention to Robert Lado's doctoral dissertation, which marked most clearly the second stage of the "scientific" period, the addition of linguistic principles to language testing. The dissertation, written by Lado at the University of Michigan under the direction of Charles C. Fries, was concerned with the construction of English achievement tests for Latin-American students. Over the next decade, Lado refined his notions of language testing, and published in 1961 a book that is a classic exposition of the structural linguist's approach to testing. It is not too much of an exaggeration to suggest that a great proportion of work in language testing since Lado (1961) is either based on it or is an attempt to answer or correct some of the points it makes.

The point of major controversy has probably been the theory of testing problems. Lado chose to set the contrastive analysis hypothesis as one of the central assumptions of his testing work, opening himself to criticism of both the general theory (cf., for example, Hamp, 1968; Di Pietro, 1971) and its application to testing (Upshur, 1962). But even this stresses the basic importance of Lado's work, for he was both insisting on and demonstrating the relevance of linguistics to language testing. He accepts completely the psychometric principles basic to testing, and explains them clearly enough for language teachers and even linguists to understand, but he leaves no doubt that linguists, with their understanding of the nature of language, must be the ones to set the specifications for language tests.

There was at the time still an easy congruence between American structuralist views of language and the psychological theories and practical needs of testers. On the theoretical side, both agreed that knowledge of language was a matter of habits; on the practical side, testers wanted, and structuralists knew how to deliver, long lists of small items which could be sampled and tested objectively. The structural linguist's view of language as essentially a matter of item-and-arrangement fell easily into the tester's notion of a set of discrete skills to be measured. It is (superficially at least) not too hard to build an objective multiple choice test from a structural grammar.

The marriage of the two fields, then, provided the basis for the flourishing of the standardized language test, with its special emphasis on what Carroll (1961) labelled the "discrete structure point" item. He noted that, "The work of Lado and other language testing specialists has correctly pointed to the desirability of testing for very specific items of language knowledge and skill judiciously sampled from the usually enormous pool of possible items. This makes for highly reliable and valid testing (Carroll, 1961)."

As a result of Lado's work, both language teachers and linguists had full access to the field of language testing. There is an important degree of modesty in his approach, for he accepts the tester's right to establish kinds of tests and methods of judging validity and reliability, even while insisting on the responsibility of the linguist to decide what is to be tested. The three major books on language testing since then (Valette, 1967; Harris, 1969; and Clark, 1972) share a great number of his assumptions. They vary, of course, in emphasis: Valette applies the principles to languages other than English, Harris aims to be concise and practical, and Clark emphasizes psychometrics, but they are all largely within the structuralist-psychometric trend.

The major achievement of this trend has probably been the production of a number of well-designed, standardized tests, such as those administered by Educational Testing Service. The Graduate Record Examinations

Advanced Tests in various languages, the MLA Foreign Language Tests for Teachers and Advanced Students, the Test of English as a Foreign Language, and the College Entrance Examination Board Achievement Tests in various languages are all good-quality tests in this tradition, widely and confidently used to measure student progress and program success. Similar tests are now available in Great Britain and parts of Europe, and the notion of objective testing and the principles on which it is based have now spread throughout the world. Of greatest importance in this development has been the possibility of tests that can be used efficiently with large numbers of subjects over a wide geographical area. The Test of English as a Foreign Language, for instance, is now given four times a year at 112 centers in the U.S. and 260 overseas.

The structural-psychometric trend has not completely overcome the objections of the traditionalists, who continue to feel that less specific measures are still of great value. They have therefore been instrumental in the development of more reliable methods of judging the more subjective kinds of performance. The first was concerned with the judgment of written proficiency. At the same time that some scholars were showing that "objective writing" tests, which usually involved multiple choice items, correlate well with other measures, others were pointing out the kinds of techniques of shorter essays, scoring guides, and multiple judgments that add reliability to subjective marking. It was shown, then, that the traditional tests, with their obvious face validity, could be improved. The second effort went into the problem of judging oral proficiency, a skill least satisfactorily handled by the objective tests. With all its importance, speech production remains the hardest to test. "The most difficult problems arise when trying to construct tests of ability to speak a language....Suffice it to say that although the ideal of a test based on free conversation is very attractive, the problems of sampling and reliable scoring are almost insoluble, unless a good deal of time and many standardized expert testers are available (Perren, 1968, p. 115)." When a good deal of time and expert testers are available, something very good can be done, as is shown by the Foreign Service Institute testing techniques described by Randall Jones (1978). The problem with these tests turns out to be not theoretical but practical: essay tests and interview tests can be made quite reliable and objective, but it is expensive to do so. The supporters of discrete item tests seem, then, to have efficiency as well as theory on their side.

There have, however, been increasingly strong attacks on their principles, associated with two trends in contemporary linguistics. The first, which I will call the language competence trend, is connected to various views of psycholinguists. It is based on a belief in such a thing as overall language proficiency, and a feeling that knowledge of a language is more than just the sum of a set of discrete parts. The second, which I will call the communicative competence trend, is connected with views of modern sociolinguists: it accepts the belief in integrative testing, but insists on the need to add a strong functional dimension to language testing.

The issue was first raised clearly by Carroll (1961). After he had described the role of discrete structure tests, he went on to argue that they fail to meet a number of basic criteria for the measurement of language knowledge. He stressed, therefore, the need for what he called an "integrative approach," where one pays attention not to specific structural or lexical items, but to the "total communicative effect of an utterance." Such an approach has several advantages: an integrative test is broader in its sampling and less likely to be tied to a particular course of training, the difficulty of the task involved is more easily

related to a subjective standard, and it focuses on the general question of how well a learner is functioning in the target language regardless of his own language background.

Carroll is thus the first to argue for what I call the "integrative-sociolinguistic" trend: he refers in this 1961 paper not just to integrative testing, but to "communicative effect" and "normal communicative situation," and others who deal with the problem (cf. Spolsky, 1968; Oller, 1973a; Jakobovits, 1969) are clearly indebted to him.

I have discussed the basic principles of the language competence, or psycholinguistic, trend in some detail in other papers (Spolsky, 1972, 1973); they are also dealt with by Oller and Ingram in this volume. Briefly, the argument goes something like this. While structural linguistic theory held that knowledge of language was a set of habits, with the consequence for testing that it is possible to select a sample of discrete items, contemporary linguistic theory emphasizes rather the creative element of language, the infinite nature of the set of possible sentences, and the incompleteness of grammars attempting to characterize knowledge of a language. This change of theoretical view challenges the linguistic validity of discrete item tests. But there is a second relevant fact about language, either derived from the statistical theory of communication or seen as part of a pragmatic grammar: this is that knowledge of a language necessarily requires the ability to function even when there is reduced redundancy, making use of what Oller calls an expectancy grammar. Two major techniques are proposed to handle this: the cloze test [Holtzman, 1967 (and Oller in this volume)], or a modified form of it (Darnell, 1968), or the dictation test with (Spolsky et al., 1968) or without (Oller, 1971) added noise. These procedures, it is argued, have the reliability of other objective measures, and their efficiency and ease of administration, and, in addition, the stronger validity provided by the theory behind them. The proposals are new, and will need considerable evidence before they are accepted; there is no mention of the cloze procedure in Lado (1961), Valette (1967), Harris (1969), or even Clark (1972); and the first three are far from endorsing dictation. But they clearly demonstrate one of the key elements in the present trend: linguists unashamedly presenting arguments, on psycholinguistic grounds alone, on the nature of language tests.

The second part of the trend is concerned with the need to test communicative competence. This trend is, in part, a reaction to the failure of Chomskyan linguistic theory to handle the full complexity of language use any better than the theories it replaced. Sociolinguistic emphases are clear in Cooper (1968) and, with a slightly different emphasis, in Jakobovits (1969); the principles are illustrated in the tests used by Cooper in the Jersey City study (Fishman et al., 1971), and those described by Spolsky et al. (1972) and by Levenston (1975). The key arguments involved in the sociolinguistic trend are twofold, one simplifying and the other complicating the process. First is the notion that knowing a language involves being able to use it in certain circumstances: whatever the specific items a speaker can control, it is his overall ability to perform with it that counts. A subject must be tested for his ability to communicate in a given situation. In a test of Navajo-English dominance, for instance, we assumed that any six-year-old who could answer the traditional Navajo question, "What is your clan" could be considered a fluent speaker of the language (Spolsky et al., 1972). The complication is, of course, allowing for the knowledge of different varieties and the ability to handle them in different circumstances.

The various approaches to testing that I have been discussing are considered in more detail in other chapters in this volume. My main pur-

pose in this introduction has been to show how it is that linguists have come to consider that language testing is well within their province. The general line of development has been something like this. Originally, testing was simply a teacher's function, although many people believed a teacher's judgment automatically improved when he changed hats and was identified as an examiner. Next, experts on testing moved into the field with their principles. It was soon shown that psychologists alone could not develop good language tests: some linguists like Lado showed that the job needed to be shared and to depend on two kinds of expertise. Finally, a group of psycholinguists and sociolinguists, with somewhat imperialistic notions, are starting to claim the field completely for themselves. Language testing, they seem to be saying, is too important to be left to language testers.

Even though this may be an exaggeration, we still need to account for the way in which language testing is such a congenial field to linguists, whereas other kinds of subject matter testing is usually left to testing experts. The answer lies, I believe, in the fact that linguists consider the question of knowledge of language to be central to their concerns. They are all the time trying to characterize in a grammar what it means to know a language: it is thus quite reasonable for some linguists to be interested in measuring knowledge of language. Upshur suggests, "Trends in second-language testing tend to follow trends in second-language teaching, and in the United States--at least in recent times--trends in second-language testing have tended to follow trends in linguistics (Upshur, 1972, p. 435)."

I believe one can state the position even more directly. As language testing has come to be a field for linguists, it has become open to direct influence from developments in linguistics. Because linguists in the last few years have been concerned with describing knowledge of a language, and, now, knowledge of language use, language testers have been able to draw on their theories for practical implications regarding how to measure such knowledge. Language testing has thus become one of the most fruitful areas in which linguistics may be applied. More, it has become one of the areas where the relevance of linguistic theory can be quickly tested in practice, and the difficulties of practice can be easily referred to theory.

Bernard Spolsky
August 1978

The Psycholinguistic Basis

Elisabeth Ingram

INTRODUCTION

There are a great many diverse methods in use for the assessment of language command. They range from translation of literary passages, writing essays, dictation, etc. to choosing the correct alternative from a set of multiple choice items, to be answered at high speed, each testing some highly specific point of phonology, syntax, or vocabulary.

The psycholinguistic bases for these practices are not at all clear. It does not help that the delimitation of the term "psycholinguistics" is not very clear either. In this article I shall first attempt to deal with some of the interpretations of the term psycholinguistics, and then try to relate these to traditional and recent testing practices.

PSYCHOLINGUISTICS

There are 2 interpretations of the term psycholinguistics (in fact, there are several, but resolving into a dichotomy of one versus all the rest). The first is highly specific, closely linked to generative linguistics, and owes its origin to Chomsky (1965, 1968). For some people this is the only acceptable use of the term. For others, like Carroll, who, I believe, first used the term in print (1953), psycholinguistics is simply a word used to cover any area of joint interest to psychologists and linguists, regardless of theoretical orientation and degree of formality.

I shall discuss first the Chomskyan interpretation. As Kuhn (1962) has pointed out, a science (or a movement within a science) consists not only of a set of theories and of techniques and methods for discovering and describing events, but also of a set of beliefs and attitudes about the proper way of regarding the whole enterprise--about the true nature of the object under study and the correct way to approach it. In transformational linguistics and psycholinguistics, these *a priori* attitudes are particularly important.

In Chomsky's view, a language consists of a non-finite set of well-formed sentences. It is the job of the linguist to describe the universals of language--the essential but abstract categories and relations which constitute linguistic deep structure--and to relate them to the structure of actual sentences in a language, as they appear on the surface. But further, for Chomsky the ultimate aim of linguistics is to contribute to the study of the human mind. A grammar must therefore not only be descriptively adequate, it must also have explanatory adequacy: it must explain the processes that underly the functioning of the "native speaker-hearer." This obviously gives psycholinguistics a very central

place. And for Chomsky, what particularly characterises the native speaker is his grammatical intuition and his language creativity. Because he has intuition, he can distinguish grammatical sentences from ungrammatical ones, and because he is creative, he can understand and produce sentences which he has never heard before. These are not theories in the strict sense. They are deeply held convictions about the essense of things--convictions which determine the areas of interest and the methods of approach.

Chomsky also has very strong convictions about the form that a theory and a description of language must take. The theory must be generative, that is, it must result in a description which is explicit. This in turn means that it must be of such a nature that a logical machine could, given the elements and the rules, generate only the well-formed or grammatically correct sentences of the language and none of the ill-formed ones.

Through very closely argued reasoning, starting from these premises (among others), Chomsky arrived at the conclusion that the desired axiomatisation can best be achieved by employing 2 sets of rules--phrase structure rules to account for the deep structure and transformational rules to link up the deep structure with the surface structure. We are here in the area of linguistic theory proper. There has been a great deal of work done within theoretical generative linguistics, and most of it has been devoted to exploring the possibilities of accounting for language in terms of these 2 rules.

Psycholinguists within the generative framework have accepted both Chomsky's presuppositions about the nature of language and language use and his specific linguistic theory. They also accept one further characterisation of native speakers--that they possess a language faculty which consists of a competence component and a performance component. When native speakers distinguish well-formed sentences from ill-formed ones, they do so by virtue of their own competence. The aim of the linguist is to describe this native-speaker competence both in terms of certain grammatical categories and in terms of phrase structure rules and transformational rules. It is these categories and rules that the native speaker must know--tacitly or intuitively--in order to function as a native speaker. This is an extremely opaque area. The native speaker must in some way have access to this knowledge--otherwise it might as well not be there--but the competence component is in no way active. It is the performance component which underlies the production and recognition of actual sentences, and which is also responsible for the errors and shortcomings of actual language use.

The acceptance of the competence/performance distinction and of the characterisation of competence in generative terms led psycholinguists to concentrate on 2 areas. The first is sentence processing: how is it that native speakers can effect the conversion between the deep structure of sentences, which is where the basic meaningful syntactic relations are given, and the surface structure of the sentences of everyday language, where the essential relations may be obscured in a variety of ways? For instance, in a sentence such as,

The tall girl I spoke to just now is German.

we know that the noun phrase the tall girl is both the subject of the matrix sentence ("the tall girl is German") and the object of the in-

serted clause ("...I spoke to just now"), but there is nothing in the sentence which overtly marks the object function. (In this particular instance it *can* be marked overtly by inserting the relative who, but that is not the point.)

By way of attacking this problem, investigators concentrated on transformationally-related sets of sentences. In some experiments they measured how long it took subjects to match up sentences which belonged together in a transformational set--John liked the old woman; The old woman was liked by John; Was the old woman liked by John; The old woman wasn't liked by John, etc.--when such sentences were scattered among ones belonging to other transformational sets (Miller, 1962). In other experiments subjects were asked to memorise sentences to see if those requiring more transformations to generate their surface structures were misremembered as ones requiring fewer transformations. Since forgetting usually involves simplification, this would indicate that transformational complexity indicates psychological complexity, and hence would support the theory of the transformational nature of competence (Mehler, 1963).

The second area is that of child language acquisition. According to Chomsky, children are born with an innate knowledge of language universals. The process of language acquisition is one of learning how to match up these innately-given universals with the surface structure of whatever language children happen to be exposed to. So the study of child language--before they have learnt to use the transformations which result in adult sentences and while they are learning--could provide empirical verification of the postulated universals.

This specific interpretation of psycholinguistics has undoubtedly been the most dominant one in recent years, both in the sense that a great deal of work has been carried out within this framework and in the sense that it has been this approach which has made the most impact on the world. Though it will no doubt continue to be very influential for some time to come, this theory, in its pristine form, is probably not now held by very many active workers in the area of psycholinguistics. There are a number of reasons for this, but it is chiefly due to the changes that are going on within generative linguistics itself.

Two fundamental aspects of Chomsky's standard theory have been vigorously challenged by some of the younger generative linguists. For instance, in Chomsky's view, the central component of language is syntax, with semantics and phonology as secondary, or "interpretative," components. McCawley (1968) and others reject this, seeing the *semantic* component as the central generative component. And if the deepest structures are semantic, then the status of a separate and distinct deep syntactic level becomes questionable and, in any case, much less important. In the last few years it has been the nature of semantic structures which has engaged attention, and while sentences, of course, continue to be studied, a great deal of both linguistic and psycholinguistic research focuses on *words*. Linguists concentrate on the autonomous semantic structure of words, while psycholinguists--and others who do not necessarily think of themselves as psycholinguists in the restricted sense--seek to account for the processes which permit the language user to perceive words, understand them, store them in memory, and retrieve them when wanted.

The other basic view which is being challenged is that language as a whole should be regarded as a self-contained system, to be described and explained without reference to any of the vagaries of actual language use

or to diversity among language users. According to this view, the meaning of sentences is fully determined by reference to language relations and language elements exclusively. The critics of this view argue that the full interpretation of sentences often depends on the listeners' "knowledge of the world," that is, on their knowledge of extra-linguistic facts. For instance, when a British doctor says, "I'm sorry, I cannot help you; you must go to your own doctor," this can be properly understood only if one knows that there is a medical rule in the United Kingdom that a doctor can treat only those patients who are registered with him or who have been sent to him by the doctor with whom they are registered.

Nobody denies that "knowledge of the world" enters into the understanding of language. The point at issue is whether it is so important that it must be incorporated into linguistic theory. The critics claim that it must be. And if it is, then, of course, language can no longer be treated as a self-contained system.

This issue has not been resolved, and the ramifications it produces extend into the psycholinguistics of the understanding of words. There is a fairly commonly-held view that the semantic structure of words is to be accounted for in terms of a partially hierarchically-ordered set of features. For instance, the word woman contains the features *animate, human, adult, female, neutral with respect to status*, etc. If one holds the view that language is an autonomous system, words have absolute meanings which are the sum of their features. And to understand words and to distinguish the meaning of one word from all the others, the language user must in some way call up the whole bundle of features which characterises each word. This again amounts to implanting the description of the linguist into the head of the speaker.

Another commonly-held, and differing, view (for instance, for words designating objects) is that semantic features are a subset of perceptual features, and that the number of semantic features that need to be invoked is variable, depending on the setting and on the intention of the speaker. Thus, a white, round block will be designated the white one if it is among round blocks of other colours, and the round one if it is among white blocks of other shapes (Olson, 1972). And when the listener tries to identify the object which the speaker designates, he has to call up only those features which, in a given situation, are sufficient to identify the object, not the whole bundle. Olson goes on to argue that the meanings of words are built up slowly in the native speaker, through an accumulation of experiences in which the various perceptually-distinct features in turn become distinguishing. (The implication is that words have partially different meanings for different people, which will not surprise any non-linguist.)

The intellectual energy which for some years converged on the development of a single set of presuppositions has now become diffused into a number of distinct, though related, approaches within the generative framework. At the same time, other approaches--some new, some temporarily submerged--are beginning to be heard again quite generally. These approaches are very diverse, but in none is language treated as a self-contained system and in all meaning is regarded as the base component. Some of these treatments are more or less directly derived from general psychological theory. For instance, Skinner's (1957) account of verbal behaviour is a straight extension of his general theory of learning. His approach is uncompromisingly communicative: the categorisa-

tion of utterances is in terms of what the speaker is seeking to achieve.*

Osgood's (1957a) psycholinguistic approach is similarly rooted in straight psychological theory, being a sophisticated version of "neo-behaviouristic" mediation theory. But unlike Skinner, who arouses a peculiarly emotional opposition among many psycholinguists, Osgood has had a considerable influence both in theoretical and practical circles. His semantic differential technique and his specific version of a semantic feature analysis has been widely used in research on bilingualism (see Jakobovits, 1970). And his cognitively-based communications model (Osgood, 1957a) was adopted as the theoretical basis of a widely-used diagnostic test of the processes underlying children's ability to use language (see p. 9).

Cognitive psychologists have always been concerned with language because of its close connection to cognitive development and cognitive structures. With the resurgence of interest in cognitive psychology in the last 15 to 20 years, the influence of men like Piaget and Bruner has been enormous. On the whole, however, cognitive psychologists have tended to treat language and language utterances as a means of studying something else, i.e. thought processes, rather than as an object of study in their own right, so the influence of these psychologists has tended to be pervasive rather than specific. But in recent years there has been a very interesting convergence between some generatively-oriented psycholinguists and others who are Piaget-oriented. Studies have appeared in which the child's comprehension and use of language has been directly related to his ability to carry out certain operations, e.g. the ability to recognise when something is *more* or *less* than something else (Donaldson and Wales, 1970), or to recognise certain logical relations, like the agent/recipient relation (Sinclair-de-Zwaart and Flavell, 1969).

Chomsky's ideas and theories have made a great impact also within main-line psychology. He has helped to reawaken the interest of psychologists in language function, and all undergraduate curricula in psychology now have a psycholinguistic component. But there has not been a revolution--in Kuhn's sense of the term--in psychology as there manifestly has been in linguistics. There are several reasons for this. One is that there are too many psychologists in the world, pursuing too many different aims. Another and more directly relevant one is that Chomskyan psycholinguistics has no learning theory. If one accepts the idea of innate knowledge, one is effectively absolved from studying the processes of learning and from trying to account for them.

In sociolinguistic circles the main dissatisfaction is with the concept of the "idealised native speaker-hearer." Hymes (1971), Labov (1969), and others do not accept as adequate the notion of the native speaker as a sentence-producing and sentence-judging machine, chugging away regardless of circumstances. They argue strongly that, in addition to accounting for how people construct sentences, it is also necessary to account for how they learn when and how to use them.

---

*I include Skinner in this survey because, while in linguistic and applied linguistic circles there is a general impression that he was killed off some time ago, there is, in fact, a very busy and lively group of people working away on operant conditioning and verbal learning, quite undeterred by fashions in other circles. And there are quite considerable spin-offs in the fields of clinical and social work.

Another recent development, specifically relevant to language testing, is the rediscovery of the virtues of pragmatics. This is a general trend (following naturally from a preoccupation with formal systems), but in testing its chief advocate is John W. Oller, Jr. (see pp. 39-57 of this volume for a discussion of his views). There is no point in setting down the arguments in favour of the pragmatic view twice, but it may be useful to examine briefly the relation between pragmatics and the various interpretations of psycholinguistics.

In the strict sense of the term, pragmatics derives from the study of formal systems. In such systems, syntax deals with the structure of expressions, semantics deals with the meaning of expressions (without reference to anything outside the system), and pragmatics is concerned with the relationships between expressions in the formal system and anything else outside of it. When Oller (1970c) states that "pragmatic facts of language are those having to do with the relations between linguistic units, speakers and extralinguistic facts (p. 99)," he is using the term more or less in its original sense. According to this definition, pragmatics is a superordinate term, covering any kinds of interdisciplinary activities in which linguistics is involved--psycholinguistics, sociolinguistics, speech and communication studies, etc. But Oller goes further; he asserts that language cannot usefully be studied as a self-contained system, that pragmatic facts must be built into the linguistic account itself. This makes claims about what linguistics should be about and obviously runs into a fair amount of opposition.

The pragmatic facts Oller is particularly concerned with are the processes of comprehension. For him, comprehension is not a matter of computing compatible interpretations of sentences in a vacuum; what is important is the expectations of the listener. Language in use is always concerned with something; listeners expect that what they hear will make sense, so they match up the incoming signals with what they know about grammar and discourse and the world. In a very informal way, this is in line with current cognitive trends in psychology, where perception and comprehension and recall is thought of as an active process where the individual matches his existing structures with the outside signals he is receiving (Neisser, 1967).

It is important to note that Oller's concern is with the comprehension of passages rather than with the interpretation of single sentences. This is the continuous concern in educational circles, since this is what a reader has to do in real life. The original competence/performance distinction was formulated in terms of well-formed sentences, considered one at a time, and the utterances that matched, or did not match, such sentences, with no built-in provision for knowledge of the world or the intentions of speakers or the expectations of listeners. To the extent that we do concern ourselves with discourse rather than with sentences, and admit the relevance of extra-linguistic factors, the applicability of the competence/performance concept, as formulated by Chomsky, becomes doubtful, if not irrelevant.

The problem of relating surface structure to deep structure remains. The difficulty is that nobody is very sure what deep structure is any more. It seems intuitively more satisfying to suppose that cognitive and relational notions such as agents and actions, assent and denial, and location in space and time are at the base of language, rather than nodally-related noun phrases and verb phrases, but we are a very long way from any kind of explicit model, or even one that is moderately

agreed to.

The current concern for comprehension relates to the view that language is to be regarded as a means of communication, both in mother tongue and second language teaching. (This is important from the standpoint of educational needs, as well.) One approach to the problem of comprehension is to try to analyse the component skills. Carroll (1972a), evaluating research carried out on a high school population, suggests that the components may be lexical knowledge, grammatical knowledge, the ability to locate "facts" in paragraphs (which presumably involves knowledge of the rules of discourse), and the ability to make inferences, i.e. to go beyond the data given. He suggests further that the first 3 may be analysed in terms of the actual language used, while the fourth would be some sort of general cognitive ability. Many would consider that this sort of approach is not psycholinguistics. It depends on one's definition of the term. If psycholinguistics must have its roots in theoretical linguistics and/or theoretical psychology, then it isn't. But if psycholinguistics is a general term covering studies into the human processing of language, then it is.

PSYCHOLINGUISTICS AND TESTING.

The essential truth about nearly all kinds of tests is that the only theory they are based on is test construction theory, which is a kind of applied statistics. Current intelligence tests are not based on any coherent or explicit cognitive theory; language tests are not based on any coherent or explicit psycholinguistic theory. Their sole justification is that they work, i.e. one can make better decisions on the basis of the information that they provide than one could without that information.

Practices in language testing are influenced by 2 things: by pre-theoretical views about the nature of language and language use and, as Upshur (1972) has pointed out, by trends in teaching practices. Language teaching practices are in turn influenced by a number of factors: by economic circumstances (for instance, during the depression in the United States, only reading skill was aimed at because the planners could not assume that students would be able to receive more than 2 years of language learning in school); by general educational trends; occasionally by linguistic, psychological, and sociological theory; but, again, perhaps most of all by convictions about the nature of language.

From time to time linguists have had great influence on language teaching. In the early part of the century, Jespersen provided a scholarly grammar of English which was also admirably suited for pedagogical purposes. Moreover, because of his views about language and about life, he was a vigorous and influential advocate of what would now be called the situational approach.

The structural linguists of the 1940's and 50's were again highly influential, both in terms of getting across their views on the nature of language and in terms of linguistic description. The view that spoken language is primary led to a considerable increase in emphasis on spoken skills, which in turn led to the construction of tests for spoken language. Three-quarters of Lado's (1961) pioneering work on language testing is devoted to the description of testing formats which deal with spoken language in some way. Curricula were drawn up in terms of lists of specific syntactic structures, selected and arranged according to the

prevailing grammatical descriptions. This fits in admirably with multiple choice testing techniques of the sort that are now sometimes referred to as discrete-point testing. It was under these influences that language testing became a flourishing business and an important part of language teaching technology.

When it comes to generative linguistics and generative psycholinguistics, the impact on classroom teaching trends has been minimal. "... an examination of recent second language textbooks shows how little of any consequence has been contributed by the theory of transformational grammar itself to the development of teaching material (Lamendella, 1969, p. 270)." This no doubt sounds paradoxical in view of the tremendous amount of debate and discussion and persuasion that has been going on. What has happened is that teachers, as usual, have been selective. Some of Chomsky's terminology and some of his views about the nature of language have been enthusiastically adopted. The notion of "creativity" has been accepted as central. The concept tends to be used in a general liberating sense, referring to the marvellous complexity of language and the untrammelled capacity of the native speaker to exploit its resources. It is not, in general, thought of as having any particular or restricting psycholinguistic or pedagogical implications. "The intuition of the native speaker" is a handy phrase for replacing the awkward "Sprachgefühl." "Competence" is popularly proclaimed to be the aim of language teaching, but it is difficult to see what practical consequences this reformulation has had. The idea of a separate and distinct "faculté de langage," not subject to the same principles of development and learning as other human capacities, has, however, caused much confusion and uncertainty (cf. Carroll, 1971). What is a teacher to do in the face of this mysterious faculty, particularly when the essential categories and relations of language are said to be innate anyway?

We are back to the absence of a psycholinguistic theory of learning. In oversimplified presentations, *habit* has become a dirty word, but nothing workable has been put in its place. Appeals have been made to *rule-based* or *cognitive* learning, but this, in Carroll's phrase, is merely "a kind of verbal overlay" which tends to add to the confusion, since it is in contradiction to the idea of the separateness of the "faculté de langage."

At the level of theory, as distinct from pre-theoretical considerations, the preoccupation with transformational rules has had only limited effects, and none in the classroom. Jacobovits (1970) was brave enough to offer a specific suggestion. Though he thought there was no theoretical basis for imitation and repetitions, he suggested that if exercises are to be given, they should provide practice in carrying out transformational conversions between different, but transformationally-related, sentences:

> From a theoretical point of view the development of grammatical competence should be facilitated by getting the learner to perform a set of transformations on families of sentences (e.g.: I cannot pay my rent because I am broke; if I weren't broke I could pay my rent; given the fact that I have no money, I cannot pay my rent; how do you think I could possibly pay my rent if I am broke; since I am broke the rent cannot be paid; to pay the rent is impossible given the fact that I have no money (p. 106).

This and other similar suggestions have failed to be convincing.

The reason why generative linguistic theory and generative psycholinguistic theory has had so little impact on language teaching practices is that teachers have, rightly or wrongly, failed to discover any relevance in it to their work. There have been no empirical consequences that are of any great practical value. There have been attempts to teach languages through some sort of transformational approach (e.g. Rutherford, 1968), but they have met with no great success; teachers and theoreticians dislike them equally. There is, however, one field in language teaching where transformational grammar could become important. It depends on whether the tentative return to being willing to give some grammatical explanation in the classroom will gain ground. An eclectic pedagogical grammar would be certain to contain transformational accounts of selected grammatical areas (see Allen and Widdowson, 1975).

Though no transformational influences have reached language testing via the teaching situation, there have been theory-motivated attempts to try out transformational-type tests directly. When Pimsleur started assembling his Language Aptitude Battery (1966), he tried out a subtest requiring skill in embedding sentences then referred to as double-based transformations. The students were presented with pairs of sentences which they had to transform on the lines of the model:

John claims it
John is right } ⟶ John claims he is right.

The subtest failed to show any correlation with success in learning foreign languages and was eliminated very early.

Similar attempts to derive tests directly from transformational theory have not proved workable. Brière (1972) incorporated a transformational subtest in the series of tests which were developed to test the proficiency of American Indian children. The children were given a simple declarative sentence and were asked to transform it into a negative, an imperative, or a question. Neither the Indian children nor the control group of native English-speaking children had much success with it. And in the administration of another subtest, consisting of having children repeat simple sentences, the investigators found no relationship between the transformational complexity of the various sentence patterns and the children's ability to repeat correctly.

In contrast, the Illinois Test of Psycholinguistic Abilities (Kirk et al., 1968), which is explicitly based on Osgood's communications model and empirically researched in the usual psychometric way, is widely used as a diagnostic tool with children showing various kinds of language or developmental deficiencies. This presumably is because people who have to make practical decisions about individual children have found that it has practical value. The test has also been used with normal children, and various subtests have been found to correlate with reading skill (Newcomer et al., 1975) and with language dominance in bilingual children (Zirkle et al., 1974). Although it is claimed that the test is theory-based, when one examines the various subtests, they seem to be based on very broadly-conceived notions of human functioning, rather than on any very specific model. This may have something to do with its usefulness in dealing with complex skills. (The subtests are concerned with memory for auditory and visual signals; the ability to check truth

value of spoken and written statements, with reference to knowledge of the world and to pictures; the ability to supply missing words in analogical frames, to describe pictures, and to show knowledge of vocabulary items; as well as the ability to deal with more formal aspects of language: spelling simple words; supplying missing letters in words; supplying the correct inflections of nouns and verbs; and choosing appropriate comparators, prepositions, and anaphoric pronouns.)

Oller's pragmatics underlies his very active and successful advocacy of the use of cloze tests for testing foreign language proficiency, as well as of the use of dictation. Both techniques represent very interesting and promising new departures in testing. This might appear to contradict the opening statement in this section that tests are not based on theory, but I regard Oller's pragmatics not as a theory in the strict sense, but as a conviction about the nature of language and language use--and such convictions do influence testing.

The emergence of cloze tests and dictation tests are 2 of the main new features of language testing. They represent not only pragmatics, but also the new interest in "global" or "integrative" techniques of testing, as distinct from discrete-point testing. This again derives from convictions about the nature of language: if language learning is to be regarded not as the mastering of a series of grammatical structures or transformational sets, but as learning how to communicate effectively, then perhaps testing should elicit "language behaviour" rather than "language-like behaviour."

Discrete-point testing has been strongly criticised in some of the recent testing literature. There have even been suggestions that such tests should be done away with altogether. In order to discuss the controversy, it is necessary to return briefly to a consideration of test theory.

Tests are interesting only when they work, that is, when they accurately measure the characteristics we wish them to measure. That brings us to the knotty problem of validity and criterion measures. How do we know that a test works? The standard answers are: because we get high correlations with total scores on test batteries made up of a number of different subtests or because we get high correlations between test scores and external estimates of the characteristic we are interested in. (Strictly speaking, this only pushes the problem one step further back, i.e. how do we know that the test battery or the external criterion is itself valid? But that is not my concern in this article.)

By way of illustrating the problem, let us consider 3 tests: (a) a sound discrimination test (Ingram, 1968); (b) a dictation test (Oller), with revised figures given by Rand (1972); and (c) a test of oral communication (Palmer, 1972). All 3 tests involve spoken language, but are otherwise very different. For the sound discrimination test, students have to match a single word recorded on tape with one of 3 written words, and this is neither integrative nor language-like. The dictation test is derived from a familiar classroom device, and is integrative but not exactly reflective of language behaviour. The test of oral communication is an extremely interesting (and recent) development foreshadowed by Upshur's (1972) discussion of the need for such tests. The examiner and the student look at strips of 4 or 5 pictures. In one version the student asks questions until he identifies the picture the examiner has in mind. In another version the student has to describe the picture he has selected until the examiner can identify it. The scoring is either

time-needed-for-identification or time-plus-error count. This test is both integrative and elicits language behaviour.

The first requirement of a new test is and always has been that it must correlate highly with the total score on a battery of assembled subtests. Most batteries contain a mixture of discrete-point and integrative subtests; for instance, grammar tests are usually discrete-point and comprehension tests are integrative. Sample correlations for the 3 tests cited above, with their respective test-battery totals, are:

	N	*Correlation test/total*
Sound discrimination	320	.85
Dictation	81	.84
Oral communication (pro-test)	33	.79

All the totals include a grammar and a comprehension subtest, apart from other types of subtests, so in all both integrative and discrete-point elements are present.

The consequence is that the only new tests which survive, whether discrete or integrative, are those which work in the same direction as the cumulative total of *both* integrative and discrete-point results. If there really is a clear-cut distinction between integrative techniques and discrete-point techniques, then by our procedures we deliberately fuzz the distinction.

A new test should also correlate with some external criterion. Because of the need to place overseas students appropriately in universities and colleges, grade point averages (GPA) are often taken as such a criterion. The correlations obtained between *any* subtest or test-total, whether integrative or discrete-point, with GPAs are extremely low. This is obviously because factors other than command of the language which is the medium of instruction enter into academic success. So GPA is not a good criterion for estimating how good a test is as a measure of language command. A better criterion is to be found in the judgement of experienced teachers. Ranking lists produced by teachers who know their classes well probably constitute the most valid criterion available, according to Vernon (1960) (provided the ranking includes only one group or class at a time). Part of the validation of the English Language Battery (Ingram, 1970) consisted of correlations between subtest and total scores and teachers' rankings. The teachers were asked to rank for all-over command of English. In the table below some of these correlations are set out for 2 discrete-point tests and 2 integrative tests. Because of the vagaries of figures relating to small numbers, I have quoted results for 3 groups, all of which were made up of young-adult students. (The ranks were converted to z-scores before they were correlated by the product-moment formula.)

Subtest	*Type*	*Correlations*		
		Group 1 *N*=37	Group 2 *N*=37	Group 3 *N*=42
Sound discrimination	discrete-point	.66	.24	.71
Listening comprehension	integrative	.68	.64	.71
Grammar	discrete-point	.68	.38	.78
Reading comprehension	integrative	.43	.70	.65

There is nothing in these figures to suggest that there is any intrinsic difference between discrete-point tests and integrative tests as regards the amount of agreement obtained between test scores and the estimates of teachers. It is, of course, true that these subtests and their items were ones which survived several stages of item analysis and correlations with totals, but that is true of any properly-researched test. It is, in any case, quite unnecessary to suppose that one has to make an either/or choice, that if one approves of integrative tests, one should therefore disapprove of discrete-point ones. This "disjunctive fallacy," as Carroll calls it, stems, it seems to me, from misunderstandings about the nature of language command.

Firstly, a test seeks to measure accurately a given characteristic in an individual drawn from a given population. The most obvious way of finding out how good a person is at doing something is to take a job sample: if you want to know how good somebody is at writing an essay, set him to write an essay. But as is well-known, due to a number of operative factors, the results may well vary: people write better essays at some times than they do at others and, just as important, judges judge better at some times than they do at others. Job samples are inherently valid, but tend to be unreliable because of this variability, thus lowering their validity. This is why multiple choice testing came into existence: multiple choice tests are highly reliable when properly constructed. But they do not necessarily possess inherent validity; in language testing they may, for instance, elicit language-like behaviour rather than language behaviour. So the validity, or lack of validity, of such tests has to be empirically demonstrated by comparision with a criterion. This, in my opinion, also holds true for job-sample tests. Once the validity has been demonstrated, however, it is immaterial what type of test we are dealing with. If a test works for the purpose it was intended to, then that is all that matters.

Secondly, though we have no great understanding of the nature of language processes, we at least know that they are very complex. It is therefore highly unlikely that any single type of test will reflect all the facets of that very intricate human faculty: language command. For any full assessment, as distinct from quick screening jobs, a number of different types of subtests are more likely to give an accurate picture than any single measure, and, within limits, the more difficult the subtests are, the greater the chances of sampling language behaviour adequately.

A PEDAGOGICAL APPROACH

Testing is an educational method. For a very small minority it is a subject of study and research in its own right, but in the wider context it is a practical tool, ancillary to teaching. Most teachers are practical people--they have to be--and on the whole they do not find highly abstract theoretical models very useful. In order to be useful, an analysis of learning--a description of language--must relate fairly straightforwardly to actual situations, to directly observable and recognisable dimensions. There is no implied criticism here of either teachers or theory-makers; I am merely stating a truism about their different preoccupations and purposes.

If testing is ancillary to teaching, then it must be accounted for on the same basis as teaching, that is, in terms of a not too abstractly-

formulated model of learning. In my view, it is not necessary or useful to set up a model for language learning which presupposes that language learning is essentially different from all other kinds of learning. But it is essential to look for an account which recognises that there are distinct, though interdependent, learning processes. I believe this to be true at any level of abstractness, but certainly, at a practical level, it is the only way of making sense of the many highly diverse teaching and testing practices which actually occur, and which actually work, when appropriately employed. The most useful account of this sort that I know of is Gagné's (1965) analysis of the conditions of learning, which is specifically aimed at educational contexts. I have elsewhere attempted to show its relevance to the second language learning and teaching situation (Ingram, 1975). I shall not repeat the arguments here, but, by definition, if the account is relevant to learning processes, it must also be relevant to testing practices (insofar as they work).

Gagné recognises a number of types of learning, all hierarchically related. The least complex and most basic type is a very simple form of perceptual learning--learning to recognise and distinguish recurring objects and events. This underlies all other forms of learning and is generally difficult to exemplify in a pure form, because most of it happens in early childhood, and other and more complex forms of learning supervene almost immediately. But the process is very clear in second language learning. For instance, in order to differentiate,

entendre; attendre [ātādr]; [atādr]

one must distinguish nasalised from non-nasalised vowels, and in order to differentiate,

cent vents; cent vins [sā vā]; [sā vɛ̃]

one must be able to tell one nasalised vowel from another.

Gagné makes it quite clear that there is no disjunction between this kind of very basic learning and the more complex forms, such as concept learning. Concept learning is essentially a matter of learning how to categorise partially different objects or events under one heading because they possess certain criterial characteristics, e.g. Alsatians and terriers and dachshunds are all dogs, or because they are functionally equivalent in some way, e.g. guns and knives and arrows are all weapons. There can be no concept learning of this sort unless perceptual learning is secure, i.e. unless we have learned to identify and distinguish objects and events in the first place. Similarly, the ability to use words appropriately depends on a series of conceptualisations: categorisations of non-linguistic objects and events must be learned; semantic categories and syntactic classes must be respected. The objects that are categorised may be more or less abstract, but ultimately all concepts that have empirical reference relate back to the perceptual world and perceptual learning.

Two other forms of learning which Gagné recognises are chaining and problem-solving. Chaining is that form of learning which enables us to produce as a smooth sequence an activity which has several component parts. For instance, verbal chaining enables us to produce and recognise a phrase or an utterance as a unit, to operate rules of agreement

and concord without computing each element separately, to get word-order right, to produce the more stereotyped utterances of social exchanges, and to fit everything into the appropriate intonation and stress contours.

In problem-solving learners have to structure a given task so that they can make decisions about relevant concepts and procedures. Whenever learners are asked to induce a given grammatical rule, they are invited to produce problem-solving behaviour. Problems can be difficult in 2 ways, either because the necessary conceptual structure is complex or, quite often, because the relevant concept just does not occur to people. It is, for instance, fairly difficult for native speakers of Germanic languages to induce unaided the rule for the use of the possessive son and sa in French: the fact that it is the concept of grammatical gender which is relevant, and the gender of the word functioning as object at that, seems initially very strange to such speakers.

I have mentioned 4 of Gagné's forms of learning--perceptual identification, chaining, concept learning, and problem-solving--and in the barest possible way indicated the links with certain language learning phenomena. Now consider them in relation to language testing formats. Perceptual learning provides a direct rationale for tests of sound discrimination, such as the one described on p. 10. Chaining can be seen to underlie those test formats which test the learner's ability to operate the obligatory rules of language, for instance, those concerning morphology and those concerning the sequencing of elements, and also those formats which test the learner's easy recognition of predictable patterns and conversational sequences. Concept learning is obviously relevant to all aspects of language use. In testing it is directly required by items which ask the learner to choose the appropriate form in light of a given context: the learner must categorise the occasion indicated by the context as being an instance of a class of occasions which call for the selection of one language form rather than another, for instance, the use of the present perfect rather than any other tense form. Finally, problem solving is an indispensable element in tests of comprehension. The learner must assemble his knowledge of grammar and vocabulary, his knowledge of the world and of the rules of discourse, to enable him to identify the facts and the conclusions and the implications of the passages he is asked to interpret.

This approach is not as elegantly simple as Chomsky's model (or Skinner's, for that matter). But it is serviceable, and that, in a teaching/testing context, is what matters.

Psychometric Considerations in Language Testing

John L.D. Clark

In the language testing context, the term "psychometrics" can be most usefully defined as any and all utilizations of numerical data and related logical operations in the service of developing, using, and interpreting the results of measurement activities carried out upon language learners or potential learners. In any given measurement activity, the psychometric procedures involved are properly dependent on the purpose which the measurement activity itself is intended to serve, and their appropriateness and adequacy are judged by the extent to which they contribute to the accomplishment of the intended purpose.

It is useful, in this regard, to define three broad categories of purpose within the language testing area. The first is *prognosis*, briefly described as the prediction of an individual's future achievements in language learning on the basis of currently available measures of a linguistic or other nature. A second measurement purpose is the *evaluation of achievement*, in which the intent is to determine the extent to which the student has learned ("acquired," "mastered," etc.) elements of linguistic content formally presented in a course or other controlled learning situation. A third broad area of measurement purpose is the *evaluation of proficiency*, that is to say, the determination of the extent to which the student is able to utilize the tested language for such real-life receptive or communicative purposes as reading magazines or novels, conversing with friends on topics of general interest, and so forth. In proficiency testing, the manner in which the measured proficiency has been acquired is not at issue: indeed, the testing process and test content should be completely independent of the student's language learning history.

In view of the extremely close relationship between the intended purpose of a given measurement instrument and the psychometric concepts and procedures appropriate to it, the discussions in the following pages have been sectioned according to the three categories of testing purpose identified. Within each section, the major concern will be to identify those aspects of psychometric practice most suited to the development, use, and interpretation of the test instruments in question, and to relate these to the format, content, and pragmatic purposes which the tests themselves are intended to serve.

PROGNOSTIC TESTING

The basic function of prognostic testing in the language learning con-

text is to use currently available information about a student to predict the level of accomplishment which he or she is likely to attain at some future time, after having followed a particular language learning program or activity. The degree to which scores on a given test or other quantifiable measure, such as rank in class or course grades, can accomplish this predictive purpose depends on the extent to which these data correlate, in a statistical sense, with achievement test scores or other quantifiable criteria of "success" used at the completion of the learning program. The correlational relationship is usually expressed by means of the Pearson product-moment correlation coefficient, which ranges from zero, indicating a complete absence of relationship between scores on the predictor and criterion measures, to 1.0, indicating a perfectly consistent relationship.[1] The higher the correlation, the more accurate the prediction, in the sense that there is a decreased statistical probability that a given prediction will be inaccurate.

The development of effective prognostic techniques is thus, in large part, an attempt to find tests or other measures which correlate highly with appropriate indices of (later) language success. Grade averages, rank in class, tests of general intelligence, and other measures generally available in student records have for many years served in the prediction of language learning success; these efforts have been reviewed by Henmon et al. (1929), Salomon (1954), and Pimsleur et al. (1962). Predictive value has also been sought through more specialized techniques, including tests of musical ability (Blickenstaff, 1963), measures of articulatory precision (E. Pike, 1959), and psychological profiles (Morgan, 1953).

Aptitude Test Development

An intensive search for effective predictors of language learning ability that could be readily and uniformly administered to prospective language students was carried out by John B. Carroll during the early 1950's in the context of the intensive foreign language courses conducted at the Army Language School in California and at other government training centers (Carroll, 1962). The research technique used was to administer test batteries consisting of a large number of experimental tasks to students entering the language learning programs, and to select, through factor analytic techniques, a much smaller number of tasks which preserved most of the predictive power of the original larger batteries. The major outcome of the Carroll studies was the publication of the *Modern Language Aptitude Test* (Carroll and Sapon, 1959), consisting of 5 separate subtests entitled Number Learning, Phonetic Script, Spelling Clues, Words in Sentences, and Paired Associates. Each of these subtests was intended to tap a competence related to the ability to learn a foreign language without requiring the student to be familiar with any language other than English.[2]

Carroll's search for higher predictor-criterion correlations (and hence, more effective prognosis) was relatively successful. In the test manual for the *MLAT*, Carroll was able to report correlations as high as .71 between *MLAT* scores and high school language course grades, compared to correlations of .35 to .52 for the Otis I.Q. test and other measures of general intelligence.

Although such results did represent an appreciable improvement in predictive power, it should be noted that a correlation of .71 accounts for only slightly more than half of the statistical variance present in the criterion scores. The remaining "unpredicted" variance reflects influences

not accounted for by student scores on the prognostic test; these may be hypothesized to include differing levels of student motivation (which, in some instances, could counter-balance a lower degree of intrinsic language learning ability), tutoring or other special study opportunities during the course of instruction, and various other factors.

Carroll Model of School Learning

A conceptual framework within which the nature and influence of these "other-than-aptitude" variables might be empirically analyzed was suggested by Carroll a number of years ago in his "model of school learning" (Carroll, 1963). According to this model, a student's success in accomplishing a given learning task can be represented as a mathematical function consisting of the following elements: the student's "aptitude" for the task in question; "ability to understand instruction," as determined by measures of overall intelligence and verbal ability; extent of "perseverance," as indicated by the amount of time the student is willing to spend in active study and presumed to reflect the level of motivation; the "time available for learning"; and the "quality of instruction" provided.

A powerful implication of such a model is the notion that students with a low level of measured aptitude for language study can, nonetheless, reach an acceptable level of accomplishment if other variables in the equation are suitably adjusted--for example, if more formal learning time is provided or if more carefully developed instructional materials are made available. These concepts may appear commonsensical to the practicing language teacher; nonetheless, their integration into the formal model proposed by Carroll is significant in that it clearly postulates the contribution to be made by each variable toward a criterion of measured achievement.

In order to validate the Carroll model and render it useful for instructional planning and prediction, it would be necessary to quantify each of the component variables for experimental study. Detailed procedures for gauging students' motivation for, and attitude toward, language study have been developed by Wallace Lambert and his associates (Gardner and Lambert, 1972; Lambert et al., 1968), and several of the scales and questionnaires used in the Lambert studies have been incorporated in the *Foreign Language Attitude Questionnaire* prepared by Leon Jakobovits on behalf of the Northeast Conference on Foreign Language Teaching (Tursi, 1970). Measures of this type could be expected to serve as indicators of "perseverance" in the Carroll model. Measures of "ability to understand instruction" are available in tests of general intelligence. "Time available for learning" could be quite easily quantified in a programmed instruction context and, notwithstanding current difficulties in accurately measuring "learning time" in the usual classroom and homework situation (Packard, 1972), effective quantification in these settings is basically a matter of improved observational and recording techniques.

The most elusive variable in the Carroll model is without doubt that of "quality of instruction." However, the use of interaction analysis procedures for classroom teaching (Moskowitz, 1970) and more precise formulations of effective teacher behaviors as judged by panels of experienced teachers (Hayes et al., 1967) provide encouraging signs

that reasonably satisfactory measures of this component may be available within the not too distant future. Multiple regression techniques and other statistical procedures are available for use with the Carroll model as soon as the necessary measures have been defined and data obtained for representative students and course combinations. The considerable practical value of a predictive system based on this model would be to permit a highly individualized prescription of the types of courses and lengths of study that students having various combinations of language aptitude, intelligence, and motivation would require in order to reach defined learning goals.

Selection of Criterion Measures

So far in the discussion, interest has been focused on the predictor measures as such, whether a single predictor--as represented by grade average, I.Q. score, or *MLAT* score--or the multiple predictors implied by the Carroll school learning model. The magnitude of any predictor-criterion correlation is also highly dependent on the nature of the criterion measure itself. It is unfortunately often the case that some readily available measure--such as the final course examination or a standardized test that happens to be on hand--will be adopted as the criterion measure for a predictive study, with little consideration of the extent to which it accurately represents the specific achievements which the prognostic measure was originally intended to predict. For example, a certain prognostic test that is intrinsically a highly accurate predictor of listening comprehension and speaking ability might show only moderate or low correlation with an end-of-course examination consisting predominantly of reading comprehension questions and writing exercises.

The proper selection of criterion measures is of special importance in large-scale research studies aimed at the experimental identification of promising predictor measures. Since the statistical procedures used in these studies operate to maximize the prediction of the criterion without regard to its nature, it is crucial that the criterion represent the most valid measure of the desired achievement available. In this respect, further advances in the area of prognostic measurement must rely, at least in part, on corresponding increases in the sophistication and precision of the criterion measures of language attainment against which they are validated.

EVALUATION OF ACHIEVEMENT

Tests used in the evaluation of achievement are focused on measuring the student's acquisition of course content--that is to say, those aspects of phonology, lexicon, and structure to which the student has been formally exposed in textbooks, classroom sessions, or through other instructional means. Within the achievement testing area, two subclassifications are possible, based on the degree of detail which the test results are intended to reflect. Tests which undertake to determine the student's acquisition, or lack of acquisition, of discrete elements of course content (for example, mastery of each of the vocabulary items introduced in a textbook unit) can be referred to as *diagnostic* achievement tests. *General* achievement tests, on the other hand, are directed at measuring the student's ability to combine several different aspects of course

content in situations which more closely approximate ordinary language use. Even though the content of a general achievement test may be more global and more "realistic" than that of a diagnostic achievement test, it continues to share the primary characteristic of all achievement tests in that it is properly based on only those combinations and recombinations of language elements that have previously figured in the formal instruction.

It should be emphasized that the *procedures* followed in developing an achievement test are uniformly applicable to any type of course or course sequence. Regardless of the theoretical or pragmatic guidelines used in the initial specification of course content (for example, contrastive analysis, functional load, situational utility, or simply the informed judgment of practicing teachers),[3] the achievement measurement question is always that of adequately representing--within the content of the test itself--the content of the instructional syllabus on which it is based.

With the possible exception of an achievement test on the first lesson of a beginning course, it would be impossible to include in any test instrument of administerable length all of the linguistic elements to which the student is exposed in the instructional setting. The specification of test content, in virtually every instance, must involve sampling, from among an extremely large number of potentially testable elements, those which can be considered to stand in for a wide number of similar elements not formally tested. Unfortunately, the identification of meaningful domains of "similar" elements is an extremely complex matter, and the more highly diagnostic the test, the more evident are the problems involved.

Problems in Diagnostic Testing

Some of these difficulties have been previously cited (Clark, 1972b), using as an example the diagnostic testing of "the written forms of the French *passé composé*." Within this general area, the proper achievement testing strategy would be to identify various content domains which could be considered homogeneous for testing purposes in the sense that student success or failure on a given item within the domain could be taken as indicative of similar performance on the other items in that domain. A proposed domain might be the different personal forms of a single specified verb (*je suis allé, tu es allé, il est allé*, etc.). Students answering a "*tu es allé* question" correctly would be expected to perform correctly on a "*je suis allé* question" or on any other component of this particular paradigm, and those failing the tested item would be expected to miss each of the other elements of the domain, were they to be tested on them.

It is obvious that the way in which the domains are specified is of crucial importance to the extrapolation of the testing results, and that the student's language learning history must also be taken into account in formulating these domains. For example, the *je suis allé, tu es allé...* domain might be appropriate for students who have, in their course work, been introduced to all personal forms of this verb. The same domain would not, however, be usable for classes in which only the "*tu*" form had been introduced at the time of testing.

For a given course of instruction, it might be feasible, although certainly arduous, to specify a number of testing domains based on care-

ful analysis of the content and sequencing of the teaching materials, and then to include each and every element of these domains in a lengthy validation test. Domains for which student testing results were not uniform across elements would be reformulated and retested; for any domains showing homogeneous results, individual elements could subsequently be drawn on a statistically random basis for inclusion in a smaller, operational test form.

Within the usual classroom situation, the prospects for such detailed test preparation activities would not seem encouraging. However, educational publishers developing new textbook programs and accompanying test materials might find this a reasonable procedure. It should also be noted that such an approach would permit the essentially simultaneous development of several alternate test forms, each having highly comparable content and measurement characteristics.[4]

A second fundamental difficulty in diagnostic achievement measurement is that of designing testing formats and individual test items which accurately and unambiguously measure the specific behaviors in question. It is unfortunate that multiple choice procedures, although admirable from the viewpoints of scoring speed and objectivity, do not lend themselves well to the diagnostic testing of discrete linguistic accomplishments. One drawback in this regard is the probability of correct response by chance. This probability is at the highest level for 2-option or "true-false" items, for which the student has a 50% chance of answering the item correctly in the absence of any knowledge of the linguistic element tested. The likelihood of successful chance response can be reduced somewhat by increasing the number of answer options per item (in 4- or 5-choice items, the chance success probability is .25 and .20, respectively), but beyond a total of 4 or 5 options per item, the item writing task becomes extremely difficult and time consuming.[5]

A second means of reducing the chance success factor is to incorporate into the test more than one item based on the same content element and to require the student to respond correctly to each of these items before mastery of the element is assumed. The statistical probability of a student's answering each of a series of multiple choice items by chance becomes very low with just a few items (for example, .008 for a sequence of three 5-choice items). Blatchford (1971) has made use of this technique in developing a diagnostically-oriented test of Chinese grammar. However, despite the statistical appeal of this procedure, the time required for test administration is appreciably increased when 2 or more items must be presented for each of the elements to be tested. A further drawback to this approach is that alert students may be able to note certain formal similarities among the various items dealing with a single element and deduce the appropriate answers solely on this basis.

In addition to the problem of chance response in using multiple choice items for highly diagnostic purposes is the difficulty of designing item stems and response options so as to eliminate the possibility that the student will be able to use information unrelated to the element tested in arriving at a correct response. For example, in a 4-choice vocabulary item presumably testing a single lexical item (the keyed answer), the student might be able to rule out the other proposed answers as inappropriate without actually understanding the meaning of the intended word.[6]

The possibility that the student may, in many cases, be able to find the correct response by following a linguistic path other than the one

intended, together with the unavoidable statistical probability of correct response on a purely fortuitous basis, render multiple choice techniques of dubious validity in testing situations which attempt to certify the student's acquisition of discrete, highly specific elements of course content. More appropriate formats for diagnostic testing purposes would include the great number of fill-in or other "constructed response" techniques presented and discussed in Lado (1961), Clark (1972a, 1975a), and Valette (1977).

If it can be assumed that the student's response to a diagnostic test item[7] is an accurate indication of his mastery (or lack of it) of the linguistic element in question, a test based on a number of such items (each dealing with a different element) may be thought of not as a single instrument for which one total score would be generated, but as a series of individual, one-item tests, with a separate ("pass-fail") score available for each item. Although such a high degree of diagnostic specificity is possible in theory, the sheer data processing and interpretation burden which the one-item, one-score principle imposes on classroom teachers and students alike would make its full-scale implementation unfeasible in most cases. For example, a 10-item diagnostic test administered to 30 students would yield 300 separate "scores." Administration of 15 such tests over the course of a school year would produce 4,500 separate items of information on student performance which would have to be tallied, reported, and interpreted.

An attempt at handling large quantities of diagnostic test data through computer techniques has been made by Poulter (1969), who used mark-sense cards as the student response medium for language laboratory quizzes. The Center for Curriculum Development (1971) at one point offered a computer-based scoring service for tests in its *Voix et Images* French program in which individual item responses were stored and retrieved for individual students and the class as a whole, together with printed statements of the linguistic aspect tested in each case. Computer-assisted test administration, scoring, and diagnostic reporting has also been undertaken by Boyle et al. (1976). With few exceptions, however, these and similar techniques are not readily available to the classroom teacher.

An alternative to single-item reporting and interpretation is the combining of several test items into broader units of content which still retain some degree of diagnostic value. For example, 2 levels of score reporting were provided for in an experimental test of "spoken Spanish grammar" developed by Educational Testing Service for use in Peace Corps language testing projects (Educational Testing Service, 1971). At the more detailed level, each response was scored separately, permitting diagnostic statements such as, "the student can produce the third person singular present tense form of *vivir* in a single sentence context using known vocabulary." At a second, more general, level of scoring, a group of related items was considered to constitute a "mini-test" for a somewhat broader category (e.g. "present tense verb forms"), with the score for each mini-test reported as the number of correctly-answered items within that category. This second technique permitted identification of areas of student strength and weakness at the level of "present tense verb forms," "possessive pronouns," "definite and indefinite articles," and so forth. Although this second-level scoring procedure did not yield the great informational detail of individual-item scoring, the data handling aspects were considerably less onerous, especially for score

information accumulated across students and test administrations.

An important conceptual and practical task in the future development of diagnostically-oriented tests of language achievement will be the adoption of a level of specificity which permits a useful degree of instructional feedback without exceeding the information processing capabilities of the students or teaching staff. Despite the problems involved in defining such a level, and in developing diagnostic achievement tests generally, these efforts should be more than repaid by the informational feedback which such instruments can provide in the service of increased student motivation (Cartier, 1972; Marso, 1969; Pack, 1972; and Steiner, 1970) and language course planning and improvement (Parent and Veidt, 1971; Valette and Disick, 1972).

General Achievement Tests

General achievement testing is by definition a less specific and less highly controlled type of evaluation in which diagnostic precision gives way to the presentation of longer and more natural language sequences. Within this framework, the use of multiple choice formats is not necessarily proscribed. Since the measurement focus is on whole-test performance, the effects of chance response are diffused over the test as a unit, and a certain proportion of the total test score is formally or implicitly discounted as attributable to chance factors. The possibility that students will take somewhat differing linguistic paths to a correct answer is also of lesser concern because of the more global measurement intent.

Although diagnostic standards may be relaxed for tests of general achievement, these tests must continue to be based on lexicon and structures to which the student has been exposed in the course of instruction. In this regard, the use of externally-prepared standardized instruments for achievement testing purposes must be conditioned on the extent to which they incorporate the lexicon, structures, and other elements of linguistic content presented in the course of instruction. To the extent that test content and instructional content differ, the usefulness of the external test as a measure of specific course achievement is diminished. Carroll (1969) and Valette (1969) independently raised this point in their discussion of testing results for the so-called "Pennsylvania study" (Smith, 1970), in which scores on the *MLA-Cooperative Foreign Language Achievement Tests* (Educational Testing Service, 1965) were used as criteria of accomplishment in "audiolingual" and "traditional" courses. Valette found that a large proportion of the vocabulary used in the MLA-Cooperative Reading Test did not appear in the textbooks used by the audiolingual classes, and on the basis of this suggested that the Cooperative Test was not a valid measure of achievement for the audiolingual group.

The problem of content in using externally-prepared tests as measures of course achievement can be obviated to some extent by careful prior examination and selection of the test instruments. Cox and Sterrett (1970) have suggested a statistical procedure in which only those test items which are judged to reflect course content would be included in calculating total test scores. This procedure would effectively remove the influence of content-inappropriate items, and, provided that the unscored items constituted only a small proportion of the total items in the test, there would be little adverse effect on test reliability.

EVALUATION OF PROFICIENCY

The purpose of proficiency testing is to determine the student's ability to use the test language effectively for "real-life purposes," that is to say, in vocational pursuits, for travel or residence abroad, or for such cultural and enjoyment purposes as reading literary works in the original text, attending motion pictures or plays in the test language, and so forth. In all cases, the measurement emphasis is on the extent to which the individual is capable of utilizing his knowledge of, and facility in, the language to accomplish some desired receptive or communicative purpose. In contrast to achievement testing, which is explicitly based on the nature and content of the student's language learning history, proficiency testing focuses entirely on the examinee's ability to perform pragmatically useful tasks in the language, without regard to the manner in which that ability was acquired.

Within the proficiency testing category, it is possible to distinguish *direct* and *indirect* procedures. From a theoretical standpoint, the most direct procedure for determining an individual's proficiency in a given language would simply be to follow that individual surreptitiously over an extended period of time, observing and judging the adequacy of performance in the language-use areas in question: buying train tickets, ordering a meal, conferring with colleagues on work-related matters, conversing with friends on topics of current interest, writing a note for the plumber, ordering business supplies by correspondence, and so forth. It is clearly impossible, or at least highly impractical, to administer a "test" of this type in the usual language learning situation. Nonetheless, the development of proficiency measurement procedures that can properly be considered "direct" must be based on approximating, to the greatest extent possible within the necessary constraints of testing time and facilities, the specific situations in which the proficiency is called upon in real life.

Validity Considerations in Proficiency Testing

The formal correspondence between the setting and operation of the testing procedure and the setting and operation of the real-life situation constitutes the face/content validity of the test--the basic psychometric touchstone for direct proficiency tests. The face/content validity of a given instrument must be determined by close examination and analysis of the testing materials and procedures themselves, and this determination is necessarily logical and judgmental, rather than statistical, in nature. This concept may be somewhat disturbing to statistically-oriented test developers and users, who might prefer some numerical index of validity, perhaps a "coefficient of face/content validity" analogous to the predictive validity coefficients associated with prognostic tests. However, since a direct proficiency test is, in effect, its own criterion, it must necessarily be evaluated by informed inspection rather than through statistical means. The judgmental nature of face/content validity should not in any way be considered a disparagement of this validation process: as succinctly expressed by Rulon (1946), "[face/content validity] sounds as though it were a rather superficial thing; as though we should require some more conclusive proof of the test's validity. Actually, there can be no more conclusive proof (p. 290)."

Direct proficiency testing can encompass the measurement of student

skill in any of the 4 language modalities. For example, the *Graduate School Foreign Language Tests*[8] provide measures of the student's proficiency in reading verbatim excerpts from journal articles and other texts appropriate to given areas of graduate school specialization. Direct measures of writing proficiency require the student to produce materials such as business letters, reports, and other documents at issue in specified real-life writing situations. For purposes of discussion, it will be useful to concentrate on an area of language proficiency of high current interest to students, teachers, and language testers: the ability to communicate orally in face-to-face language situations.

As has been emphasized by a number of authors (Cooper, 1970; Jakobovits, 1969, 1970; Paulston, 1974; Spolsky, 1968; and Upshur, 1972), an individual's ability to communicate effectively in a given language cannot be considered directly proportional to his mastery (or lack of it) of specified lexical items, grammatical structures, or other discrete elements of performance. As a consequence, instead of using linguistic inventories as a point of departure for setting test specifications, the developer of a communicative proficiency test must be concerned with arranging testing situations that are the closest possible facsimiles of real-life communication situations. Instead of evaluating the linguistic accuracy *per se* of the examinee's performance, the tester must concentrate on determining the extent to which the examinee is able to convey various types of information in an accurate, efficient, and situationally appropriate way.

With respect to appropriate settings for a direct test of communicative proficiency, the presence of a live interlocutor is probably indispensable for adequate face/content validity. Published speaking tests using tape-recorded stimuli to which the student replies are some steps removed from a real communicative situation in that they do not allow for the speaker interaction and instantaneous alteration of content characteristic of face-to-face conversation. Except for telephone conversations and other communication-at-a-distance situations, real-life oral communication also involves facial, gestural, and other visual cues which cannot be provided in a test situation except on a face-to-face basis.

However, the mere fact of a face-to-face conversation is not of itself a sufficient demonstration of validity: close attention must also be paid to the topical content of the conversation and to the psychological/interpersonal relationships that are established during the course of the test. It is probably futile to hope that the affective components of a formal testing situation will ever closely approach those of the real-life situations which the test attempts to reflect. As Perren (1967) expresses it: "...both participants know perfectly well that it is a test and not a tea-party, and both are subject to psychological tensions, and what is more important, to linguistic constraints of style and register thought appropriate to the occasion by both participants." Nonetheless, for the sake of test validity, every effort must be made to minimize the "examination" aspects of the conversation in favor of a more natural and encouraging ambiance.

Scoring Procedures

In addition to the validity of the test setting and administration procedure, there is also the question of validity of the scoring procedures used. The degree of scoring validity depends on the extent to which the

scoring system represents examiner judgments of the student's ability to convey information in an efficient and situationally appropriate way, rather than the grammatical accuracy, correctness of vocabulary, or other linguistically-oriented aspects of the student's performance. It is, however, often difficult to separate "communication" and "linguistic accuracy" in scoring practice. For example, the scoring criteria for "level 2" performance on the Foreign Service Institute's language proficiency interview (Rice, 1959) are reproduced below, with emphasis (italicizing) added to indicate those portions involving judgments of linguistic accuracy rather than of communicative performance as such:

> Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e., topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; *accent*, though *often quite faulty*, is intelligible; *can usually handle elementary constructions quite accurately* but *does not have thorough or confident control of the grammar*.

A similar intermingling of communicative and linguistic criteria is seen in the description of "elementary speaking proficiency" on the *English Proficiency Chart* of the National Association for Foreign Student Affairs (again, emphasis added):[9]

> Asks and answers questions on daily personal needs and familiar topics with *very limited vocabulary*; makes *frequent basic errors in structure and pronunciation*.

In addition to the validity question for scoring procedures is the problem of scoring reliability. Although scoring reliability is not a significant problem in testing procedures based on multiple choice or short response formats, it assumes substantial proportions in situations where human judges must assign numerical ratings to longer and less highly structured samples of language behavior. Two types of scoring reliability are at issue in such instances: intra-rater reliability, which refers to the extent to which a given rater is able, repetitively, to assign the same score to a given test performance, and inter-rater reliability, which refers to the extent to which 2 or more raters assign the same score to a given performance. Low intra-rater reliability is a serious problem, since it indicates that the standards of scoring judgment are not stable even among individual judges. Low inter-rater reliability is also a troublesome matter in that the obtained score becomes dependent in large part on the rater involved: examinees who happen to draw a more lenient rater stand to benefit by comparison with examinees whose performance is evaluated by a more severe rater.

Studies of intra- and inter-rater scoring performance have primarily been conducted in the area of written production or "essay testing" in the examinee's native language, as comprehensively reviewed by Coffman (1971). Tests of oral communication in either native or second languages pose technical problems in that it is substantially more difficult to "represent" the student's performance for repetitive scoring by the original

rater or other independent raters (Clark, 1975b); this situation has doubtless contributed to the general paucity of scoring reliability studies in the speaking proficiency area. As reported in the test manual, an inter-rater reliability study of 100 speaking test tapes in the *MLA-Cooperative Test* battery yielded a 2-rater correlation of .59, using a scoring procedure based primarily on judgments of linguistic accuracy rather than on overall communicative ability. In a small-scale reliability study of the FSI interview carried out at Educational Testing Service, the scores of 2 raters simultaneously present at 80 interview sessions coincided as to basic score level (on a 6-point scale) in approximately 95% of the cases. Notwithstanding the information provided by occasional limited studies of this type, the scoring reliabilities of direct proficiency interviews (and of tests of speaking ability in general) remain to be comprehensively investigated and documented.

Several procedures might be suggested to increase the scoring reliability of direct tests of communicative proficiency. For example, the interviewers could be asked to cover specified topical areas or to ask a fixed series of questions of each examinee. Or various scoring aids could also be devised, such as the "Factors in Speaking Proficiency" chart developed by FSI staff for use along with the original verbal ratings of competence (Wilds, 1975). This chart breaks down the student's performance into the categories of pronunciation, grammar, vocabulary, fluency, and comprehension. Weighted scores are assigned to each category, and the student's final rating is derived from a total score across categories which is then reconverted to the original verbal scale.

Although a high degree of scoring reliability for direct proficiency tests is certainly an important goal, such reliability should not be sought at the expense of face/content validity. In the first example above, detailed advance structuring of the content and sequencing of the interview would make it less representative of the often digressive conversations typical of real-life communication. In the second example, the compartmentalization of the student's responses into different linguistic categories for scoring purposes, together with the assignment of fixed score weightings to each category, could be expected to do some violence to the final rating as a direct measure of communicative proficiency. Although it may be hoped that testing procedures can eventually be developed which combine a high degree of face/content validity with high scoring reliability, in cases of conflict, considerations of validity should take precedence as the *sine qua non* of direct proficiency testing.

Indirect Proficiency Tests

Indirect proficiency tests are also intended to assess the extent to which the student is able to function appropriately in real-life language-use situations. However, unlike direct proficiency tests, indirect measures are not required to reflect authentic language-use contexts and, indeed, they may in many cases bear little formal resemblance to linguistic situations that the student would encounter in real life. One example of an indirect proficiency measure is the "reduced redundancy" test developed by Bernard Spolsky (Gaies et al., 1977; Spolsky, 1972; and Spolsky et al., 1968). In this procedure, the student is asked to listen to and transcribe a series of sentences in the test language which are accompanied by a specified degree of electronically produced background noise. Development of this test is based on the theory that individuals

who have a high level of global proficiency in the language will be better able to utilize the reduced number of linguistic cues available in acoustically distorted speech and will thus be able to perceive the recorded stimuli accurately at lower signal/noise levels. Although a case might be made for the existence of a limited number of "reduced redundancy" situations in real-life contexts (for example, telephone reception over faulty equipment), the Spolsky technique does not, in general, reflect the kinds of language-use situations in which the student would be expected to operate in real life.

The usefulness of the Spolsky test and other indirect proficiency measures does not, however, depend on the tests' face/content validity but on the extent to which the test scores are found to correlate, on a statistical basis, with more direct measures of the proficiency in question, simultaneously administered during the test validation phase. The practical utility of these concurrent validity studies and the obtained correlations lies in the extent to which it thereby becomes possible to predict, on the basis of an examinee's indirect test score alone, the score that he or she would be expected to obtain on the more direct measure when the latter cannot be administered for reasons of cost or complexity of administration.

An indirect testing technique which has received considerable recent attention is the so-called "cloze" procedure, in which the examinee attempts to replace words previously deleted from a continuous text. Originated by W. L. Taylor (1953) in connection with native language learning, the cloze procedure was intensively examined by Carroll et al. (1959) as a possible measure of foreign language proficiency within the College Entrance Examination Board testing program. In the Carroll study, only moderate reliabilities were found for French and German cloze tests based on an every 10th-word deletion pattern, and their operational use in the College Board program was not recommended. More recently, a number of additional investigations have been carried out using various adaptations of the original cloze procedure. Darnell (1968) developed a "clozentropy" procedure in which the test responses of native speakers were used to define and weight acceptable answers according to an information theory model. A 200-item test using the clozentropy technique was found to correlate to the extent of .84 with scores on the *Test of English as a Foreign Language (TOEFL)* for a group of 48 non-native speakers of English studying at the University of Colorado. A major disadvantage of the Darnell approach is the need for computer assistance in the complex scoring procedure.

In other studies, Oller and Inal (1971) administered an English cloze test in which only prepositions were deleted and obtained a correlation of .75 with total scores on the UCLA English placement examination, consisting of multiple choice and free response questions covering vocabulary, grammar, reading comprehension, and dictation. Oller (1972b) found that a scoring system which gave credit for any contextually acceptable word (not necessarily the original deletion) resulted in higher test reliability than the exact-word-replacement method, as well as a higher (.83 vs. .75) correlation with the UCLA placement examination. Superiority of the "any-acceptable-word" scoring procedure was not, however, corroborated in a later study by Hanzeli (1977). Recent experimentation has also been conducted using the cloze procedure in a multiple choice format (Jonz, 1976; Griffin et al., 1978).[10]

The correlational results so far obtained with indirect proficiency

measures, especially the cloze procedure, allow for some optimism that techniques can be developed to estimate, in an efficient and economical manner, the student's acquisition of "real-life" proficiencies that are directly measurable only through more elaborate and more expensive techniques. However, some cautionary observations should also be made, as follows:

•With a few exceptions (Pike, 1973; Hinofotis, 1976), experimental studies involving indirect procedures have used as comparison measures multiple choice tests or other instruments that do not in themselves have a high degree of face/content validity as direct measures of the language proficiencies in question. As has been frequently urged in the testing literature (Carroll et al., 1959; Clark, 1972b, 1975b; Lado, 1960; and Spolsky, 1968), it would be very desirable to carry out detailed studies in which direct language interviews and other highly face-valid techniques would serve as the criteria against which perception-in-noise tests, cloze tests of various types, and other experimental measures could be correlated and operationally compared. In addition to permitting close examination and comparison of proposed indirect measures, such investigations would focus attention on, and quite probably bring improvements to, the direct measurement techniques themselves.

•The magnitude of the correlation between indirect and direct proficiency measures may be affected by the specific language learning history of the individuals tested. Although high correlations between a printed cloze test and a direct measure of oral proficiency might be obtained for examinee groups whose language experience has included routine contact with both spoken and written materials, the same relationship might not be shown for examinee groups having other learning backgrounds. In this regard, the oral proficiency level of students whose language training has been largely restricted to either the spoken mode (as in some Peace Corps training situations) or the written mode (as in reading-knowledge-only courses) might be under- or over-predicted, respectively, using correlational results obtained from student groups with more heterogeneous language backgrounds. Additional investigation of the influence of diverse learning histories on indirect test performance would appear indicated, as well as more global studies of the interrelationships among language modalities as they affect test performance generally.

•The "representational value" of indirect proficiency tests is much less than that of direct proficiency tests. Whereas the amount and quality of language accomplishment represented by a given level of performance on a direct proficiency test is readily apparent to the examinee and other interested persons, the same cannot be said of indirect test scores. When used in the classroom and other instructional situations, indirect proficiency testing should properly be accompanied by explanatory materials which permit an appropriate extrapolation of the test results to specified types and levels of real-life performance.

SUMMARY

The discussions in the preceding sections have dealt with the major psychometric considerations at issue in 3 types of language testing activities. Prognostic measurement involves the use of test instruments or other available measures to determine, through correlational techniques, the level of language accomplishment that specified students would be expected to attain if they were to follow particular instruc-

tional programs. The *MLAT* and similar language aptitude tests can contribute to the effective measurement of at least one component of language learning success, but a potentially more effective prognostic technique involves a "systems approach" based on the Carroll model of school learning, in which several student-related and instruction-related variables are considered simultaneously in estimating, for a given student, the probable outcome of a number of alternative learning programs and strategies. Improvement in the quality of the criterion tests used to define language learning success is also considered of great importance in the continuing development of prognostic techniques.

In the area of achievement testing--defined as the measurement of the student's acquisition of course content--a major psychometric concern is that of appropriately sampling that content within the confines of an administratively-feasible test. A suggested empirical approach to the content question is the establishment of operationally homogeneous content domains from which individual elements can be drawn for testing. Diagnostically-oriented achievement tests attempt to certify the student's acquisition (or lack of it) of discrete, minutely specified content elements. Multiple choice techniques are not considered well-suited to diagnostic testing activities because of statistical and logical factors which render single-item data ambiguous as to the student's mastery of the point ostensibly tested. Completion exercises and other techniques requiring actual language production are considered more appropriate for highly diagnostic testing, even though they lack the scoring speed and convenience of the multiple choice format. Practical difficulties in handling the large amounts of data generated by diagnostic testing at its most highly specific level may make it necessary to combine tested elements into somewhat larger units for scoring and interpretation purposes.

General achievement testing is a more global type of measurement in which larger and more natural linguistic units can legitimately be presented or elicited. The basic requirement in achievement testing--that the test content be derived exclusively from course content--is applicable to general achievement tests as well as to diagnostic tests: in this respect, the appropriateness of using externally-prepared instruments for general achievement testing depends in large part on the extent to which the external instruments adequately mirror the content of the language program in which they are to be used.

Proficiency testing involves measuring the student's ability to utilize the tested language for pragmatically useful purposes within a real-life context. Direct proficiency testing, discussed primarily in terms of the testing of face-to-face communicative proficiency, represents an attempt to duplicate the real-life language-use situation as closely as possible within the test setting, as a demonstration of high face/content validity. Scoring procedures for direct proficiency tests must demonstrate real communicative criteria and a high level of both intra- and inter-rater reliability.

Indirect proficiency tests have the same measurement purpose as direct proficiency tests but derive validity as proficiency measures through a correlational relationship with direct proficiency tests, rather than through the face/content validity of the instruments themselves. "Reduced redundancy" tests and various types of cloze procedures are examples of indirect measures which show considerable promise as indices of language proficiency in situations where more direct tests cannot be administered. However, direct proficiency tests will continue

to be needed, both for administration in their own right wherever possible and as criteria against which the adequacy and accuracy of more indirect procedures can be established.

FOOTNOTES

[1]Negative correlations ranging from zero to -1.0 are also possible; these usually occur when the scoring scale for one of the correlated measures is reversed so that better performance is represented by lower scores. Negative correlations have just as much predictive value as the corresponding positive values.

[2]Subsequent to the publication of the *MLAT*, a *Language Aptitude Battery*, based on generally similar principles, had been made available (Pimsleur, 1966).

[3]For useful discussions of alternative procedures in defining course content, see George (1962) and Perren (1971).

[4]For a fairly technical but highly useful further discussion of domain-based achievement testing, see Shoemaker (1975).

[5]Various "correction-for-guessing" procedures have been developed in an attempt to minimize the statistical effect of random guessing on multiple choice test scores. However, these involve adjustments of the student's total test score and in no way counteract the possibility of correct answering by chance at the individual-item level. Test instructions which caution the student not to guess and warn of a penalty for wrong answers may be taken at face value by some students but disregarded by others more willing to risk an incorrect response. For a review of the literature in these areas, see Diamond and Evans (1973).

[6]For additional discussion of this point, see Clark (1965).

[7]"Item" is used here in the general sense of "test question"--including short answer, completion, and other question types--and not multiple choice questions *per se*, which are seen to have drawbacks for diagnostic testing.

[8]Published as an ongoing series by Educational Testing Service, Princeton, New Jersey.

[9]See National Association for Foreign Student Affairs (1971).

[10]An extensive bibliography on cloze testing in English-second-language applications is provided by Oller (1975b).

The Sociolinguistic Foundations of Language Testing

Joshua A. Fishman & Robert L. Cooper

It is the purpose of this article to illustrate the usefulness of a sociolinguistic approach to the construction of language assessment procedures. This approach insists upon the specification of the communicative contexts in which the behavior to be assessed occurs, and can be justified on two grounds. First, language behavior and behavior toward language vary as a function of communicative context. Thus, global, uncontextualized measures of language proficiency, language usage, and language attitude may mask important systematic differences. Second, language assessment procedures have been successfully contextualized so as to gather data reflecting systematic sociolinguistic variation. This article presents evidence to support both justifications for the sociolinguistic contextualization of language assessment procedures.

SOCIOLINGUISTIC VARIATION

Sociolinguistics describes a loosely-associated set of inquiries which have as a common concern the relationships between linguistic and other social variables. (For a collection of essays reviewing the field, see Fishman, 1971.) While investigators who attempt to describe and explain these interrelationships differ in their orientations, all agree on at least one point: there are no single-style speakers and no single-style speech communities. That is to say, no one speaks in the same way all of the time, and no community is composed of speakers who all have identical verbal resources at their command. Thus, a person speaks differently when shouting at an umpire at Yankee Stadium than when delivering a formal lecture. Similarly, not all the people who "speak the same language" have the same opportunities for using it. Whereas most New Yorkers may have the opportunity to go to Yankee Stadium and to learn how to display their grievances at an umpire's call, fewer New Yorkers have the opportunity (and far fewer the desire) to learn how to deliver formal lectures.

Several examples can be cited to support this notion. Labov (1966), for example, demonstrated that in the speech of New Yorkers the presence of final or preconsonantal /r/ in words such as *car* and *park* is systematically related to the social class of the speaker and to the carefulness of his speech. Thus, New Yorkers belonging to the upper end of the socioeconomic continuum produce this sound more often than do New Yorkers from the lower end of the continuum, and all New Yorkers pronounce it more frequently when they speak carefully than when they speak casually and spontaneously.

Similarly, Fischer (1958) found that variation in the pronunciation of the present participle *-ing* by children of a New England village was sytematically related to contextual and personal variables. The variant *-in'* (as in *huntin'* and *fishin'*) was more likely to be produced in relaxed than in formal situations, i.e. in verbs like *hunting* and *fishing*

than in verbs like *reading* and *writing*, by boys more often than by girls, and by "typical" boys more often than by a "model" boy. He found a slight tendency for children from less economically advantaged families to use *-in'* more often than children from more favorable economic circumstances, but the community was relatively homogeneous with respect to social class.

Variation in the use of American terms of address has been shown to be systematically related to differential power relations and to degree of intimacy between speakers (Brown and Ford, 1961). Thus, for example, two Americans are more likely to use each other's first names when talking to each other if they have a close relationship and to use mutual title plus last name (e.g. *Mr. Smith*) if they do not. Non-reciprocal use of the first name is more likely to be found in situations of unequal power relations (e.g. employer-employee), with the more powerful addressing the less powerful by the latter's first name and receiving title plus last name from him. Also, the more powerful is typically the one who initiates a change from non-reciprocal to reciprocal use of the first name.

Another example of variation in American English can be seen in baby talk (speech addressed to infants), which is marked by a small set of lexical items, many of which involve reduplication (e.g. *choo-choo*, *bow-wow*) and special intonational contours and syntactic features (Ferguson, 1964). Americans consider baby talk appropriate for use with babies, pets, and lovers, but many, particularly men, feel embarrassed at using it in public. That Americans talk to infants (and are often tireless in their attempts to elicit speech from them) is itself culturally determined. Other groups, the Luo of Kenya, for example, believe it inappropriate to try to elicit speech from infants (Blount, 1972).

The rich collection of speech events in which inner-city American Black adolescent boys are skilled (Labov et al., 1968) presents another example of sociolinguistic variation. These events include ritual insults (playing the dozens), the recitation of epic poems (jokes or toasts), and the formal display of occult (heavy) knowledge (rifting). Each of these is stylistically marked. Rifting, for example, requires a high-flown rhetorical style and employs many learned, Latinate words, and its syntax is closer to that of Standard English than is the syntax of other speech events.

All of the above examples illustrate the sociolinguistic universal originally asserted--that there are no single-style speakers or single-style speech communities. Social and contextual variables are reflected by differences in phonology, morphophonemics, lexicon, and syntax. They are also reflected by differences in what is said. Membership in a speech community is marked not only by a shared language or language variety, but also by shared rules for its use. Thus, members of a speech community share not only linguistic competence, the ability to understand and produce the theoretically infinite set of sentences comprising the language, but also communicative competence, the knowledge of when to speak and when to remain silent, and what to say to whom and when (Hymes, 1972). Thus, for example, one of the first things American students of Hindi or Marathi want to learn to say in those languages is *please* and *thank you* because of the importance of these terms in American English (Apte, 1974). They are among the first terms which American parents try to teach their children, and they are the terms used constantly by adults. As Apte has shown, however, the use of gratitude expressions is culturally

determined. If an American wants to speak Hindi or Marathi appropriately, he must learn that in the Hindi and Marathi speech communities there are some communicative contexts in which expressions of gratitude are obligatory, there are others in which they are optional, and there are still others in which they are taboo. He must learn when to express gratitude and, equally important, when not to express it. And he must learn this as part of learning these languages' rules of speaking, which will enable him to communicate appropriately as well as grammatically.

It is not difficult to demonstrate the systematic variation that exists in the same speaker's language usage or the systematic differences that exist between the language usages of different groups of speakers. If the obvious has been belabored, it has been because such variation is typically ignored in the construction of language assessment devices. Most writers of such devices appear to view language as a monolithic entity and to have adopted the simplifying assumption of an ideal speaker-hearer who speaks the same way all of the time and does so within a linguistically homogeneous community. Such an assumption is justified when the language proficiency, language usage, or language attitude to be assessed is contextually invariant. Certainly, there are invariant behaviors which are worth assessing. For example, we may want to predict the speed and accuracy with which a person can translate articles in psychological journals from his mother tongue into a given target language. Or we may want to predict the degree to which a university student will be able to comprehend lectures in his field when the lectures are delivered in a given language. Yet even these examples are not illustrative of completely invariant behaviors. Articles in particular psychological journals or on given psychological topics may require somewhat different skills than other articles, and lectures given by particular instructors or on particular topics may require somewhat different skills than other lectures. But if the contextual variability of the behavior we wish to assess is relatively small, we may be justified in making the simplifying assumption of invariance.

Whether or not the assumption of contextual invariance is justified in a particular case, the assumption is typically an unexamined one. It is the point of this article that language assessment procedures can be improved if the assumption of contextual invariance (or variance) is made explicit. This can be done by specifying the communicative contexts for which the language behavior or behavior toward language is being assessed. If there is only one context for which the assessment is necessary or if the assessment is for substantially similar contexts, the procedure can be designed with that context or set of contexts in mind. If there are several contexts for which assessment is necessary and if these contexts have substantially different communicative requirements, the procedure can be designed to reflect this sociolinguistic variation.

EXAMPLES OF CONTEXTUALIZED LANGUAGE ASSESSMENT MEASURES

Although most language assessment devices appear to have been written on the implicit assumption that the behavior to be assessed is monolithic and contextually invariant, a beginning has been made in constructing contextualized assessment devices. Examples follow for the measurement of language proficiency, language usage, and language attitude.

Language Proficiency

Two language proficiency devices developed in connection with the description of bilingualism among Puerto Ricans in New York City (Fishman et al., 1971) illustrate the usefulness of a contextualized approach to the measurement of language proficiency. One procedure was a word-naming task administered in English and in Spanish. Respondents were asked to name, in one minute, as many different words referring to a specified context as they could. This was done in each language for each of five contextual domains: family, neighborhood, religion, education, and work. For the domain of family, respondents were asked to name as many words as they could that named things which could be seen or found in a kitchen; for neighborhood, things seen or found in a neighborhood; for religion, things seen or found in a church; for education, subjects taught in school; and for work, the names of jobs, occupations, or professions. Responses were elicited for all five domains in one language, followed by all five domains in the other. The language in which responses were first elicited was randomly chosen for each respondent. The order of domain, however, was kept constant, i.e. family, neighborhood, religion, education, and work. Directions were of the pattern, "Tell me as many English (Spanish) words as you can that name things you can see or find in a kitchen--your kitchen or any other kitchen. Words like *salt* (*sal*), *spoon* (*cuchara*), *rice* (*arroz*)." The task was individually administered to 38 adults.

When the average number of Spanish words produced was compared to the average number of English words produced, when summed across all five domains, no difference was found. On this basis the respondents could be called "balanced" bilinguals since their total, global performance was the same in each language. However, differences were observed between the average English and Spanish scores obtained for several domains. For example, more Spanish than English words were named for the contexts of family and religion. To describe the performance of the group as a whole would be misleading, however, for significant subgroup differences were observed when the respondents were divided by age and length of residence on the mainland of the United States. These subgroups, like the group as a whole, appeared "balanced" in terms of their total English and Spanish scores. However, differences were observed between the subgroups in the pattern of their relative language proficiency as exhibited by domain. For example, school-aged respondents who had received their formal education via the medium of English showed a significantly higher education score in English than in Spanish, whereas the school-aged respondents who had received their education via both languages showed no significant difference between their average language scores for that domain. Thus, the word-naming task revealed important proficiency differences associated with different interactional domains, and these differences could be explained in terms of the respondents' differential use of English and Spanish in their everyday life. These proficiency differences, however, would have been completely hidden if only a global, undifferentiated measure had been used.

It might be objected that the word-naming task is a relatively indirect measure of proficiency (although correlations between Spanish-English word-naming difference scores for particular domains, on the one hand, and more direct proficiency measures on the other were typically about .50, a respectable relationship considering the usual relative

unreliability of difference scores and the brevity of the procedure which elicited them). A more direct language proficiency test--a measure of listening comprehension in English and in Spanish--was employed with the same adults who participated in the word-naming task. It differed from conventional listening comprehension tests in that it was designed to assess comprehension in terms of specific social contexts.

The listening comprehension test's stimulus material consisted of five tape-recorded conversations between Spanish-English bilinguals living in New York. The participants in all but one of the conversations were Puerto Rican college students who spoke fluent, native English and who were adept at switching between languages. In one conversation, one of the speakers was a parish priest, who played himself in that role, and whose Spanish was fluent but not native.

Each conversation was obtained in the following manner. First, the "actors" agreed upon a social situation in which switching between English and Spanish would be appropriate among Puerto Ricans in New York. Second, they mapped out a story-line which determined the general direction of the conversation in that situation (i.e. who would say what to whom), but no scripts were prepared. The actors then assigned the roles to one another and "role played," or ad-libbed, the scene, using English when they felt English was appropriate and Spanish when they felt Spanish was appropriate. Finally, they played back the conversation to themselves to determine whether or not it sounded natural. If parts of the conversation struck them as unnatural, those portions were re-recorded and at a later time spliced into the tape. Each completed conversation lasted between two and three minutes. Transcripts of the conversations can be found in Fishman et al. (1971, pp. 675-694).

Each of the five conversations was intended to represent a different type of social context. Consequently, the relationships between speakers (e.g. mother-daughter, priest-parishioner), the locales or settings (e.g. home, rectory), the topics of conversation (e.g. the annual Puerto Rican parade, the health of an uncle), and the purpose of the interaction (e.g. offering an invitation, dictating a letter) all varied from conversation to conversation.

After a conversation had been played twice to the respondent, he was asked a series of questions designed to assess his comprehension of the passage. In addition to questions asked in order to test comprehension of the English and Spanish portions of each conversation, other questions were asked in order to assess the respondent's interpretation of various aspects of the social situation represented by the conversation as a whole. For example, respondents were asked to identify the role-relationships between speakers (e.g. boss-secretary), the degree of social distance or intimacy between speakers, the motivation underlying certain remarks made by the speakers, and, for some conversations, the educational and occupational status of the speakers.

For each conversation, the percentage of items assessing comprehension of the English portion which each respondent correctly answered was subtracted from the percentage which he correctly answered of items assessing comprehension of the Spanish portion. The percentage correct of the other types of items--assessing interpretation of various components of the conversation as a whole, such as the role-relationship between speakers--was also computed. Correctness was scored in terms of the impression intended by the actors in their formulation of the social situation to be presented.

The usefulness of contextualizing the listening comprehension test can be seen from the differences which were observed according to conversation. For example, there was a greater difference between the average English and Spanish comprehension scores for the conversations which took place within the context of work than for the conversations which took place within the context of a home. Similarly, the relationship between the ability to understand the manifest content of the conversation (what was said) and the ability to interpret the latent content of the conversation (what was meant) differed according to conversation. For example, respondents correctly answered a greater proportion of latent content items than of manifest content items for a conversation taking place within a home, whereas the reverse was true for a conversation taking place within an office. Thus, knowing what was said did not necessarily enable listeners to absorb the full communicative impact of a conversation; conversely, missing the details of manifest content did not necessarily prevent listeners from grasping the speakers' intent.

Techniques such as the contextualized listening comprehension test can be used to assess the degree to which a community's rules of speaking have been internalized. Thus, two contrasting groups, to whom the test was also administered (Anglo high school students studying Spanish and South Americans studying at a university in New York) often agreed with the Puerto Rican respondents about what was said but disagreed with them (and with each other) about what was meant. Clearly, the communicative (as distinguished from narrowly linguistic) competence of the three groups differed not only from one another but from context to context, and many of these differences would have been lost by conventional, noncontextualized measures of language proficiency.

Language Usage

Devices which are designed to assess the relative frequency with which a person employs his languages or language varieties may be misleading if they yield a single, overall score of language usage. Thus, if a person is asked what language he uses every day and if he uses one language for most everyday purposes but reserves another language for use in specified social contexts, his response that he mainly uses the first language, while true, is also misleading since it obscures his sytematic use of another language.

Two procedures for assessing language usage illustrate the advantage of obtaining information about usage in different contexts. One measure is a language-usage questionnaire developed in connection with the study of Puerto Rican bilingualism mentioned above. Thirty-four schoolchildren, aged 6 to 12, were individually interviewed. The children were asked a series of questions to assess the degree to which they used Spanish relative to English with various bilingual interlocutors in school, at church, in the neighborhood, and at home. For example, the children were asked to indicate the extent to which they used Spanish with other Puerto Rican bilingual children when playing outside on the street near their home. Responses were scored on a 5-point scale, with the exclusive use of Spanish at one end of the scale and the exclusive use of English at the other. An average rating for the use of Spanish across various interlocutors was computed for each respondent and for each context. The children reported that they used more Spanish than English in the contexts of neighborhood and family, and more English than Spanish in the contexts

of school and church. Their overall use of English and Spanish, however, summed across all four contexts, was about the same. Thus, a question designed to assess only their global use of Spanish relative to English, without reference to the contexts of language usage, would have been misleading.

Whereas the first example of a contextualized measure of language usage was obtained from self-reports, the second example was obtained from the reports of outside observers. As part of a study of the status of English in Israel, the use of English, on a busy shopping street in Jerusalem and in the shops that line it, was described (Rosenbaum et al., 1977). A transaction-count procedure (Bender et al., 1972) was employed by which the number of persons heard speaking English was determined. It was found that approximately 14% of all the persons heard (N=936) were speaking English (the majority of speakers, of course, used Hebrew). However, almost all of the interactions involving English were between pedestrians talking to each other on the street or between customers talking to each other in the shops. There was very little English observed between customers and shopkeepers inside the shops, but this was not due to the fact that the customers and shopkeepers did not know English. In fact, most of them were able to conduct transactions in that language. English was used mainly between native speakers of English, not as a lingua franca, i.e. as a medium of communication between persons who do not share the same mother tongue. In Israel, Hebrew is the lingua franca par excellence. It is the language which is expected for use between Israelis who do not share the same first language. Since almost none of the shopkeepers spoke English natively, the native speakers of English used Hebrew with them. Thus, the transaction-count procedure recorded an important systematic difference in language usage by taking into account the relationships between speakers. If only a single count had been made--of all speakers across all contexts--this difference would have been missed.

Language Attitude

Just as global measures of language proficiency and language usage may be misleading, so global measures of language attitude may obscure important systematic differences. Attitudes toward a language or attitudes toward a referent for which language serves as a symbol may vary according to the context in which the language is used. The effectiveness of a contextualized approach to the study of language attitudes can be illustrated by two studies, the first by Carranza and Ryan (1975) and the second by El-Dash and Tucker (1975).

Carranza and Ryan asked bilingual Anglo and Mexican American high school students to rate speakers of English and Spanish on the basis of voice cues alone. Following the work of Lambert (1967) and his associates, a comparison of evaluative reactions to speakers of two languages was used as an indirect measure of interethnic attitudes. Such a procedure typically employs, as stimulus material, tape recordings of speakers reading aloud a standard passage. This procedure has been criticized on the grounds that differences in ratings of speakers in the two languages may occur if the passage read represents a context inappropriate to one of the languages (Agheyisi and Fishman, 1970). Accordingly, Carranza and Ryan used two speech contexts. In one, a mother is talking as she prepares breakfast for her family; in the other, a teacher is giving a history

lesson to her class. These contexts were designed to represent home and school, respectively. Each context was recorded in each language, yielding four passages in all. Respondents were asked to rate each of sixteen different speakers (each reading one of the four passages so that each passage was heard four times) on each of 15 semantic-differential scales. Their responses demonstrated a striking interaction between language and context. The English versions of the school context were more highly rated than the Spanish versions, whereas the reverse was true for the home context. For these respondents, attitudes towards each language (or towards the group represented by the language) was in part a function of the appropriateness of the context in which each was used. If only one context had been employed, the results would have been misleading.

In the research reported by El-Dash and Tucker, the views of Egyptians toward Classical Arabic, Colloquial Arabic, and three varieties of English (American, British, and Egyptian) were studied. Respondents were asked to rate various personal characteristics of speakers heard on tape recordings, which represented each of the five language varieties. They were also asked to rate each speech variety heard with respect to its suitability for each of five contexts (at home, at school, at work, on radio and television, and for formal and religious speeches). While the respondents tended to rate speakers heard using Classical Arabic more favorably than speakers heard using the other speech varieties, Classical Arabic was not considered suitable for use at home. In this context, Colloquial Arabic was preferred. Thus, attitudes towards language use were found to be a function of communicative context. Again, a global measure would have obscured this result.

SUMMARY

This article has justified a sociolinguistic approach to the construction of language assessment devices on the grounds that it is both necessary and possible to use such an approach. It is necessary because language behavior and behavior toward language vary as a function of communicative context. That it is possible has been demonstrated by the examples of contextualized measures presented here. Authors of language assessment procedures, therefore, should consider explicitly the contexts in which the behaviors they wish to describe take place. The more the criterion behavior varies as a function of context, the more important it is to construct techniques which can reflect variation.

Pragmatics and Language Testing[1]

John W. Oller, Jr.

It is interesting that we often speak of the coinage of terms. This metaphor is doubly effective if you consider that terms are more or less *stamped* into existence by a *mentor* (if I may be allowed a bad pun), and they have a certain purchase value like any other coin. The coin metaphor is particularly apropos to the term "pragmatics" which, in the words of William James, emphasizes the "cash value" of linguistic elements as negotiable items in communication. It comes from a brand of American philosophy initiated by Charles S. Pierce at about the turn of the century, and his thinking was extended by William James, John Dewey, and Charles Morris.[2] Although all of these scholars were Americans, the methods and assumptions of what may be called a "pragmatic approach to language study" are by no means unique to Americans, nor are they a recent development.

This paper discusses in historical perspective the major concepts of pragmatics and relates them to language testing. Grammar is viewed as a theory of language competence and is characterized as a pragmatically-generated expectancy device. It is claimed that in order to adequately measure language skills, language tests--whether for first, second, or foreign language learners--must activate the internalized expectancy grammar of the learner. Empirical data showing remarkably high correlations on very different tests of language skills are explained on the basis of the postulated expectancy grammar. It is hypothesized that other tests of language skill which fail to produce high correlation with effective "integrative tests" (the term is from Carroll, 1961) are probably invalid as measures of language proficiency.

The claim that a person who learns a language internalizes a grammar, i.e. a generative system that specifies relationships between sound and meaning in the language, is now widely accepted, though there is still a lot of debate about the form of such a grammar. For instance, there is disagreement about whether it is primarily motivated by syntactic, semantic, or pragmatic considerations, whether it is more or less generated by principles of learning such as induction and substitution (Oller, 1972c), or whether it is in substantial portion already present in the brain of an infant at birth. In this connection, it may be useful to note that the transformational generative approach seems to be evolving in the direction of a pragmatically-motivated theory of grammar. The first stage of the Chomskyan paradigm was the position that syntax and semantics were independent (see Chomsky, 1957, reviewed and criticized by Reichling, 1961; Jakobson, 1959; and others); the second was that syntax and semantics were not independent but together were independent of pragmatic considerations (see Chomsky, 1965; Katz and Fodor, 1963; and Katz and Postal, 1964; also, see criticisms by Uhlenbeck, 1967; and Oller et al., 1969); the third stage which now seems to be emerging is that syntactic, semantic, and pragmatic factors are intricately interrelated and may, in

fact, be inseparable. This latest development is illustrated in the following excerpt from *Language and Mind* (Chomsky, 1972):

> It is not clear at all that it is possible to distinguish sharply between the contribution of grammar to the determination of meaning, and the contribution of so called 'pragmatic considerations', questions of fact and belief and context of utterance. It is perhaps worth mentioning that rather similar questions can be raised about the notion 'phonetic representation'. Although the latter is one of the best established and least controversial notions of linguistic theory, we can, nevertheless, raise the question whether or not it is a legitimate abstraction, whether a deeper understanding of the use of language might not show that factors that go beyond grammatical structure enter into the determination of perceptual representation and physical form in an inextricable fashion, and cannot be separated without distortion, from the formal rules that interpret surface structure as phonetic form (p. 111).

Apparently, Chomsky now sees theories of both sound and meaning as susceptible to reinterpretation due to pragmatic facts. Concerning phonetic representations, Dennis Sales and I had advanced the same basic argument as early as 1969, on the basis of a series of demonstrations showing that controlled variations in extralinguistic contexts systematically brought about changes in the stress patterns of the surface forms of utterances.

Further support for Chomsky's somewhat cautious hint is provided by D. K. Oller and Eilers (1975), who showed that the quality of phonetic transcriptions of children's utterances is improved when the transcribers are either able to guess or are told the meanings of the utterances. Even more recently, and perhaps more confidently than Chomsky, Fillmore (1973) has made a strong case for the importance of pragmatic factors in language teaching. In fact, his remarks parallel closely some of the observations on the importance of pragmatics in several earlier publications on the same topic (compare Oller, 1970b, 1971b).

The chief argument in favor of a pragmatic approach to language derives from the principle of non-summativity. A theory of language that attempts to divorce syntax from semantics can no more hope to explain language communication than a good book on spelling can hope to explain the logic of a novel. The same sort of reasoning suggests that any attempt to account for meaning apart from situational context (i.e. semantics divorced from pragmatics) is doomed to inadequacy. With the present re-examination of the whole question of the relation between language and extralinguistic contexts, it seems that a growing interest in pragmatics is likely to be a major theme in linguistic analysis for some years. Although there are many important unanswered questions, there no longer seems to be any substantial disagreement concerning the fundamental need for a pragmatically-based account of language use and language learning. Scarcely anyone is still seriously maintaining that grammar can be regarded as a self-contained entity independent of extralinguistic contexts. This seems to be an indication of progress. However, the turn to pragmatics does not represent the birth of a new approach so much as a return to a useful tradition of language study. On the other hand, it does constitute a significant change in current trends of research in the language sciences.

HISTORICAL PERSPECTIVE

As Solomon said, "There is nothing new under the sun," and much of what seems to be progress is without doubt merely a restatement in current terminology of notions that were held true by the ancients and have only been rediscovered in their modern contexts. A serious student of the nature of human communication and mental behavior can ill-afford a contemptuous attitude toward early thinking on these topics (Chomsky, 1972, and Cherry, 1965). One could compile a great compendium of pithy quotations showing that ancient scholars and many of their progeny were aware of the importance of the fact that language often relates to things other than language. What is remarkable is that some linguists in recent decades seem, temporarily at least, to have forgotten and even actively neglected so important and obvious a fact. This, of course, is the reason that it is necessary to stress the pragmatic nature of language.

One of the early indications of concern among language theorists for the pragmatic aspects of natural languages was the theorizing of the Danish Modistae in the thirteenth and fourteenth centuries. Bursill-Hall (1971) claims that they "constructed their theory [of grammar] on extralinguistic facts based on the structure of reality...(p. 35, footnote 84)." Their concern for universal grammar was a tradition continued by the Port-Royal grammarians of the seventeenth century (Chomsky, 1966; Aarsleff, 1967). The emphasis of both schools on the rational explanation of grammatical categories in terms of intrinsic logic and extralinguistic fact is evidence of their concern for what has been termed pragmatic mappings (Oller, 1975a). John Locke (1690), who was a late contemporary of the Port-Royal grammarians, even went so far as to argue that "all words are taken from the operations of sensible things,... (cited in Kuhlwein, 1971, p. 53)."

Pragmatics distinguishes two basic levels of communication that are employed to relate linguistic elements and extralinguistic situations. James Harris (1751) proposed the basis for this distinction only a century after the heyday of the Port-Royal school (and I doubt that he was the first to notice it): "The Truth is, that every Medium thro' which we exhibit any thing to another's Contemplation, is either derived from *Natural Attributes*, and then it is an *Imitation*; or else from *Accidents quite arbitrary*, and then it is a *Symbol* (his italics, cited in Kuhlwein, 1971, p. 69)." In other words, we may either use pictures or abstract symbols to map or portray extralinguistic facts. In natural languages we usually use both, though we tend to rely more heavily on abstract symbolic means for the communication of cognitive context and on facial expression and tone of voice for the conveyance of attitudinal information. As James Beattie (1788) put it, "the Natural signs of thought are those changes in complexion, eyes, features, and attitude, and those peculiar tones of voice which all men know to be significant of certain passions and sentiments (cited in Kuhlwein, 1971, p. 97)." The brandished fist may be a picture of a threatened slug in the mouth, or a symbol of brotherly solidarity, just as a smile is often a sign of friendliness, or sometimes hideous spite. But, postural and gestural changes are in themselves quite inadequate to code much of the information that the human mind negotiates. Beattie says, "when compared with the endless variety of our ideas, these Natural Signs will appear to be but few. And many thoughts there are in the mind of every man, which produce no sensible alteration in the body (p. 97)." He goes on, "Arti-

ficial Signs, or Language, have, therefore, been employed for the purpose of communicating thought, and are found so convenient as to have superseded in a great measure,...the use of the Natural (p. 97)." Whether or not we agree with Beattie's intimations about language evolving from a need to communicate more abstract notions or ideas, his remarks clearly differentiate the two fundamental modes of communication distinguished in 20th century pragmatic theories of language (Watzlawick et al., 1967).

Of course, pragmatic mappings are abstract, and simple explanations that try to relate words and things directly are quite unsatisfactory. Paralleling the earlier statement quoted from Locke, Dugald Stewart (1810) wrote, "I have...remarked the disposition of the Mind to have recourse to metaphors borrowed from the Material World.... This analogical reference to the Material World adds greatly to the difficulty of analyzing with philosophical rigour, the various faculties and principles of our nature, yet it cannot be denied, that it facilitates, to a wonderful degree, the mutual communications of mankind concerning them,... (cited in Kuhlwein, 1971, pp. 100-101)."

Though serious scholars have often noted the difficulty of incorporating into theories of language abstractions that relate words to things (Bloomfield, 1933; Harris, 1951; and Chomsky, 1957, 1965; just to mention a few), the fact that such relations exist is one that is ignored at great peril to the theories. Surprisingly, Bloomfieldian and early Chomskyan writings are about the only prominent sources of language theories that use the ostrich approach to pragmatic data, excepting possibly the positivistic philosophy of Rudolf Carnop and his followers. The lectures of de Saussure, about 1912, compiled by his students (1959), and the writings of B. Malinowski (1935), L. Hjelmslev (1954), J. R. Firth (1957), the Prague School (Vachek, 1966), Sidney Lamb (1966, 1973), M. A. K. Halliday (1961, 1977), and many others have maintained cognizance of the fact that language is used for purposes other than putting words together in a neat arrangement with other words. Since the writings of most of these more recent authors are better known and more readily available, we will just refer briefly to a note about Firth. Robins (1963) says, "meaning, the object of all linguistic analysis in Firth's approach, is function in a context, whether the extralinguistic context of situation or the intralinguistic contexts of grammar, phonology, or other subsidiary levels (reprinted in Kuhlwein, 1971, p. 9)."

Nor have linguists (excluding the Bloomfieldians and early Chomskyans) been the only scholars concerned with pragmatic aspects of language structure. Psychologists (especially Osgood, 1957b), philosophers (especially Russell, 1940), logicians (especially Reichenbach, 1947), communication experts (Cherry, 1965), and even physicists (Einstein, 1951) have persistently evidenced concern for the fact that language is intrinsically structured for the codification of information that is largely non-linguistic. Not long ago, a group of logicians and philosophers met in Jerusalem to discuss the importance of pragmatics to theories of natural languages (Bar-Hillel, 1971). More recently a journal has been created on the topic. Obviously, a great deal more could be, and perhaps should be said in this vein, but the main theme of this volume, which is language testing, draws us in another direction.

As Upshur (1972) observed, trends in language teaching have tended to trail along in the wake of linguistic theories, and trends in testing, at least in second or foreign language testing, have been similarly tagging along behind the prevalent methods and theories of language

teaching. Witness the influential writings of Lado (1957) on methods of language teaching patterned after the "scientific" analysis of language, and his companion volume on language testing (Lado, 1961). These and many other books and articles on language testing and teaching were heavily influenced by theories that deliberately ignored the data of pragmatics (Oller, 1971a, 1973a, and forthcoming). The "discrete point" method of teaching and testing are both naive concerning the fact that a totality is greater than just a heap of unrelated parts.

Fortunately, many applied linguists rejected the naivete that was characteristic of dominant theories in the late 1940's, the 1950's, and through the mid-1960's. For example, as early as 1904, Otto Jespersen, the Danish linguist, was arguing that materials designed to teach a foreign language should have *meaningful sequence* throughout. He realized that if linguistic structures are presented outside of a meaningful context, learning will be more difficult. This has subsequently been demonstrated many times over. (For a review of the literature, see Oller, 1971b.) As Jespersen put it,

> ...we ought to learn a language through sensible communications; there must be (and this as far as possible from the very first day) a certain connection in the thoughts communicated in the new language...one cannot say anything with mere lists of words. Indeed not even disconnected sentences ought to be used....When people say that instruction in languages ought to be a kind of mental gymnastics, I do not know if one of the things they have in mind is...sudden and violent leaps from one range of ideas to another (p. 11).

The reason that pragmatically based language teaching materials can be expected to be more effective has been made clear in numerous psycholinguistic studies in recent years. If the learner is made aware of the pragmatic contexts to which language structures relate, he has a much more powerful basis for subconsciously constructing and thereby internalizing the grammar of the language. The partially predictable sequence of events in communicative contexts is one sort of data that the learner can capitalize on to great benefit. In fact, if pragmatic mappings of utterances onto contexts are not made available to the learner, there is no reason to suppose that language acquisition can occur at all.

TOWARD A DESCRIPTION OF A PRAGMATIC EXPECTANCY GRAMMAR

Let us now turn our attention to the characterization of grammar as a model of underlying language competence. (Later we will relate these considerations to language testing.) We will first discuss some empirical facts of language use which suggest that one of the important characteristics of such a grammar must be a capability to generate expectancies based on contextual dependencies. With this in mind, we will use the term "expectancy grammar." Some empirical data will be cited in support of this notion, and a partial formalism will be described. Then we will consider some findings of research in language testing and attempt to draw some inferences about *valid* language tests.

For some years now, it has been popular to speak of the perception of language as a process of analysis-by-synthesis. The evidence that such a process underlies the perception of linguistic sequences--whether in

listening, reading, or some combination of the two, as when following a text that is being read aloud by someone else--is pervasive. The motor theory of speech perception, proposed by Liberman (1957) and others at Haskins Laboratories, maintained that the perception of distinctive sound segments was mediated by the articulatory processes necessary to produce those segments. This notion provided the seminal basis for the later theory of analysis-by-synthesis developed by Stevens (1960). It is well-known that some phonemes of English, for instance, and especially the distinctive intonational contours of English, are often indistinguishable without reference to higher-level contexts (Lieberman, 1967; Stevens, 1960). We have already noted Chomsky's remarks (1972, p. 111) in this vein and the evidence in support of them. Much recent research in the perception of spoken and written forms of language suggests that there is a close relationship between perceptual processes and the pragmatic structure of language (see the articles in Horton and Jenkins, 1971; also, Kavanagh and Mattingly, 1972).

It seems that the perception of linguistic sequences is mediated by an expectancy grammar that is continually formulating, modifying, and reformulating hypotheses about the underlying structure and meaning of input signals. These hypotheses are related via pragmatic mappings to extralinguistic contexts and are instrumental in the analysis of the surface form. Chomsky and Halle (1968) suggest that, "the hypothesis will...be accepted if it is not too radically at variance with the acoustic material." Or, putting it differently, the perceiver seems to rapidly alternate between a *synthesis* that is "fast" and "crude" and an *analysis* that is "deliberate, attentive,...and sequential" (Neisser, 1967).

A growing body of experimental data on listening and reading processes lends credence to the analysis-by-synthesis model of perception. As Levin and Kaplan (1971) put it, "listeners and readers alike, appear to decode sentences not only by interpreting as they hear or read, but also by anticipating what is likely to come next (p. 2)." Kolers (1971) argues that the reader or listener seldom makes a specific guess as to what word, phrase, or sentence is likely to follow. Rather, he generates a kind of readiness for a range of possibilities. It is as though the perceiver were prepared for answers to certain questions, or even that he were searching for answers to specific questions.

The synthesis, or hypothesis, that the perceiver generates as a match for the input signal can be characterized in a natural way in terms of a grammar of expectancy. The perceiver's hypothesis about the input signal is largely based on the contextual constraints that are utilized by the internalized grammar. The hypothesis that is eventually accepted is what the perceiver hears, reads, or understands. In perception, the expectations generated by the grammar are subject to modification whenever they fail to produce a sufficient match for the incoming signal. Creative errors in reading and listening provide dramatic evidence for this process. For example, one foreign student taking a dictation test at UCLA transformed an entire paragraph on "brain cells" into a fairly readable text on "brand sales." The student's rendition was similar phonetically to the original passage on a phrase by phrase basis, but completely obliterated the original content. Less remarkable examples illustrate the same underlying process. On another dictation test, for instance, students wrote "scientist's imaginations" and 'scientist's examinations" for "scientists from many nations." The latter kinds of

errors are actually quite common among non-native speakers. Richards (1971a, 1971b) has drawn some interesting conclusions about underlying processes based on learner's errors, as has Corder (1971).

Although the process of analysis-by-synthesis has been quite widely accepted by researchers as a plausible basis for the perception of language, to my knowledge the proposal that an analogous process underlies the production of language (Oller, 1973b, 1974a and b, and forthcoming) has only been tentatively suggested. Nevertheless, it seems that language production is a kind of synthesis-by-analysis. The speaker or writer has an idea that he wants to communicate, but, as Colin Cherry (1965) has said, he never really has it until he "jumps on it with both verbal feet." It seems that the speaker or writer has a notion of what he wants to say--a sort of hypothesis or prior synthesis, if you like--and he analyzes it by putting it into words. In this way, we may conveniently explain the potent observation by Dewey (1926) that the words which come out of a person's mouth often surprise that person more than anyone else. In the case of language production in contrast to perception, expectancies are the governing factor, and the physical signal is adjusted to match them. In performing the synthesis-by-analysis, it is frequently the case that details and relationships previously unavailable on a conscious level do become available consciously, and, hence, the surprise value to ourselves of the things that we say.

At this point, an empirical example may help to make clear how a grammar of expectancy serves to explain crucial aspects of language use. The example will also provide a bridge to the description of the tentative formalism for a grammar of expectancy which is discussed below.

Clark (1966) has shown that phrases referring to actor, action, and recipient in transitive sentences are differentially constrained in actives and passives. In the passive, as Levin and Kaplan (1971) observe, "the latter part...[referring to] the...[action] and the actor is highly constrained by the former part, the...[recipient]; this was not true for the corresponding parts of active sentences (p. 4)." Clark (1966) and Roberts (1966) have shown further that recall for actives and passives is governed by the uncertainties predicted by Clark's earlier experiment. The experiments of Levin and Kaplan (1971) themselves, with eye-voice-span (EVS), confirmed their prediction that EVS would increase in the middle of passive sentences but not in actives. Results that support similar generalizations have been achieved with right and left imbedded sentences. Levin and Kaplan (1971) suggest that

> the important point is that the constraints facilitate processing only insofar as they lead to the formation of successful anticipations. The reader then can test his hypothesis for himself. If it is confirmed, the previously assigned interpretation is accepted and the material can be easily and efficiently processed. If he cannot confirm his previously assigned interpretation he must backtrack and reassign interpretations, which seems easier to do in reading than in listening. To elaborate, successful hypothesis generation depends on the ability to formulate or assign some tentative interpretation to what has been read or heard (p. 13).

Further confirmation is provided by Wanat and Levin (1968) and Wanat (1974).

The empirical data supporting the operation of an expectancy grammar,

a device that generates and confirms hypotheses, is overwhelmingly affirmative. The major question that remains is what specific form a grammar of expectancy might take. In other words, how can the notion be formalized in a suggestive and helpful way. More importantly, one may want the formalism itself to be vulnerable to empirical tests in a variety of ways. In particular, it should mirror the utilization of contextual constraints in a way that naturally accounts for the data of language use. It should also be conveniently modifiable wherever it fails to predict the data.

The first attempt at such a formalism was notably unsuccessful. It was, in fact, a left-to-right finite state grammar which operated only on constraints from one word to the next. A significant inadequacy of such a mechanism was its inability to suspend processing of a certain segment while dealing with an intervening one: it had no way of remembering to return to the earlier segment in order to continue its work there. Related to this debilitating difficulty was the lack of capability in handling recursive functions where elements of indefinite length might be strung together or imbedded in complicated ways. The observations of Lashley (1951) showed the total inadequacy of such devices as models of even the simplest sorts of human behavior. The work of Chomsky (1956) provided a mathematical proof for the inadequacy of finite state models.

A second stage in the attempt to come to grips with the sequential nature of language processing was achieved independently and nearly simultaneously by a number of researchers in diverse areas. "Finite state devices," "transition network grammars," or "directed graph models," as they were variously called, were modified in important ways to achieve recursive generative capacity (Newcomb, 1963; Johnson, 1965; Conway, 1964). While in most cases the generative power of the mechanisms achieved at this stage was equivalent only to context-free phrase structure grammars, they maintained the virtue of a fairly straightforward account of the sequential nature of a great deal of language processing. This virtue is not shared by the phrase structure grammars usually written by linguists operating in the Chomskyan tradition. Moreover, recursive transition network models allow for the expression of certain grammatical regularities in more economical and natural ways than phrase structure rewrite rules. They are more convenient in that the effects of modifications in the grammar are often more readily apparent than in phrase structure grammars.

Nevertheless, without further modification in the direction of greater complexity, recursive transition models are inadequate in a number of important ways. As Woods (1970) observed, they are not able to "move fragments of the sentence around (so that their positions in deep structure are different from those in the surface structure), to copy and delete fragments of sentence structure, and to make...actions on constituents generally dependent on the context in which those constituents occur (p. 592)." Woods proposed a solution to these problems in the form of what he calls "an augmented recursive transition network grammar (p. 591)." The grammar Woods has developed does not utilize pragmatic constraints of extralinguistic context, but in spite of this limitation it provides a useful formalism for the notion "expectancy grammar" as we have used the term here. Moreover, such a grammar can be modified to incorporate pragmatic information.

The changes that Woods (1970, 1972) imposes on the "recursive transition network" in order to achieve what he calls an "augmented recursive transition network," or simply an "augmented transition network," are

various "conditions" which must be met if a transition is to be followed, as well as certain "actions" or formal operations on constituents that are to be executed if the transition is followed. The augmented grammar that Woods has developed is capable of keeping track of tentative decisions already made, and of modifying them as more input is analyzed. At the same time, it is constantly keeping track of the limits of subsequent possibilities by anticipatory operations based on the information available to a given point. Woods says, "structure building actions associated with the arcs of the grammar network allow for the re-ordering, re-structuring, and copying of constituents necessary to produce deep structure representations of the type normally obtained from a transformational analysis, and conditions on arcs allow for a powerful selectivity which can rule out meaningless analyses and take advantage of semantic information to guide the parsing (1970, p. 591)."

To illustrate the functioning of an "augmented transition network grammar," or an "expectancy grammar," we may refer to the differential processing of active and passive sentences noted earlier in this paper. Examples *a* and *b* below are parallel in several respects. However, *a* is "passive" while *b* is "active" (at least as these terms are defined in the research mentioned earlier). To be more correct, technically *a* uses a transitive verb in the passive while *b* uses an intransitive verb.

(a) The little boy was bucked off by the spotted pony.
(b) The little boy was gone away by the next day.

An expectancy grammar with the properties described by Woods (1970, 1972), with modifications to take pragmatic information into account (as do real speaker-hearers), might recognize and interpret *a* and *b* in roughly the way described below. For the sake of the example, we assume a phonetic analyzer plus an expectancy grammar. Many grammatical details are omitted, and we reify the grammar, that is we assume that it is somehow internalized in the brain of the speaker-hearer.

The first input word made available to the expectancy grammar from *a* is the. This element can be segmented from little and the following elements inasmuch as the grammar has no lexical entry corresponding to phonetic sequences of [θəl], [θəlɪ], [θəlɪr̆l̩], etc. It recognizes the as a determiner. Hence, it knows that the is the beginning of a noun phrase (of course, it assumes that the speaker has not made a false start, or any one of a number of other possibilities which the grammar would eventually rule out anyway on the basis of subsequent information). The structure building operation which is executed activates the declarative sentence routine pictured in Figure 1.

The first arc in that routine corresponds to a noun phrase sub-routine which is expanded in Figure 2. For the subject of the declarative sentence, the grammar stores in its memory register the fact that the first word in the string being processed is the determiner the. Provided this analysis is correct so far, the grammar knows that several subsequent possibilities are likely. The determiner may be followed by an intensifier or string of them, modifying an adjective or string of them, modifying a head noun. Or, it may simply be followed by a head noun. In terms of the sub-routine in Figure 2, the grammar has already taken the transition labeled "Det" to arrive at state q_1. The grammar anticipates a noun to follow which is represented in the transition to state q_6.

In scanning the next word it discovers little. This phonetic sequence

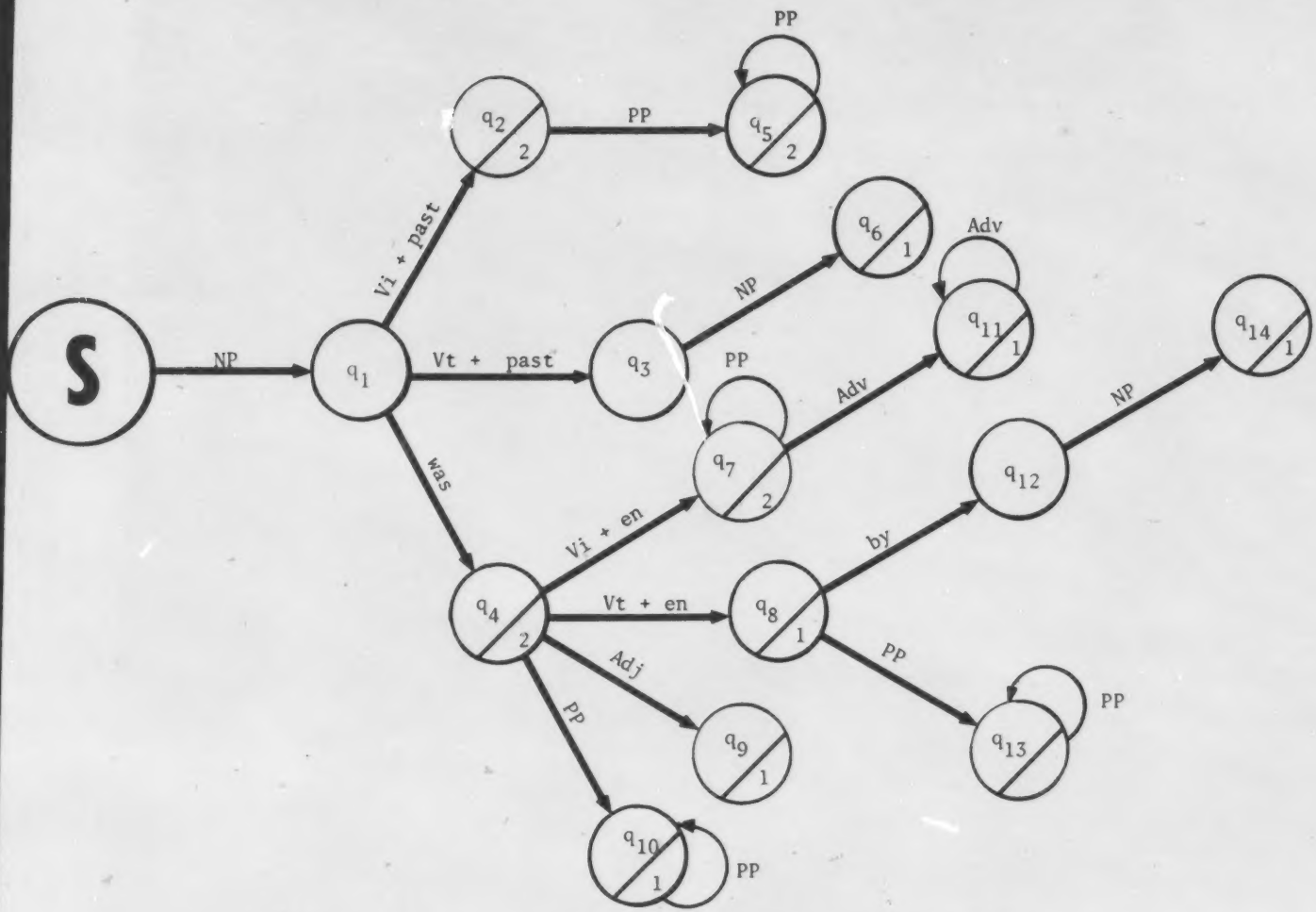


Figure 1. A Fragment of a Recursive Transition Network Grammar: A Routine for Generating Some Declarative Sentences in English.

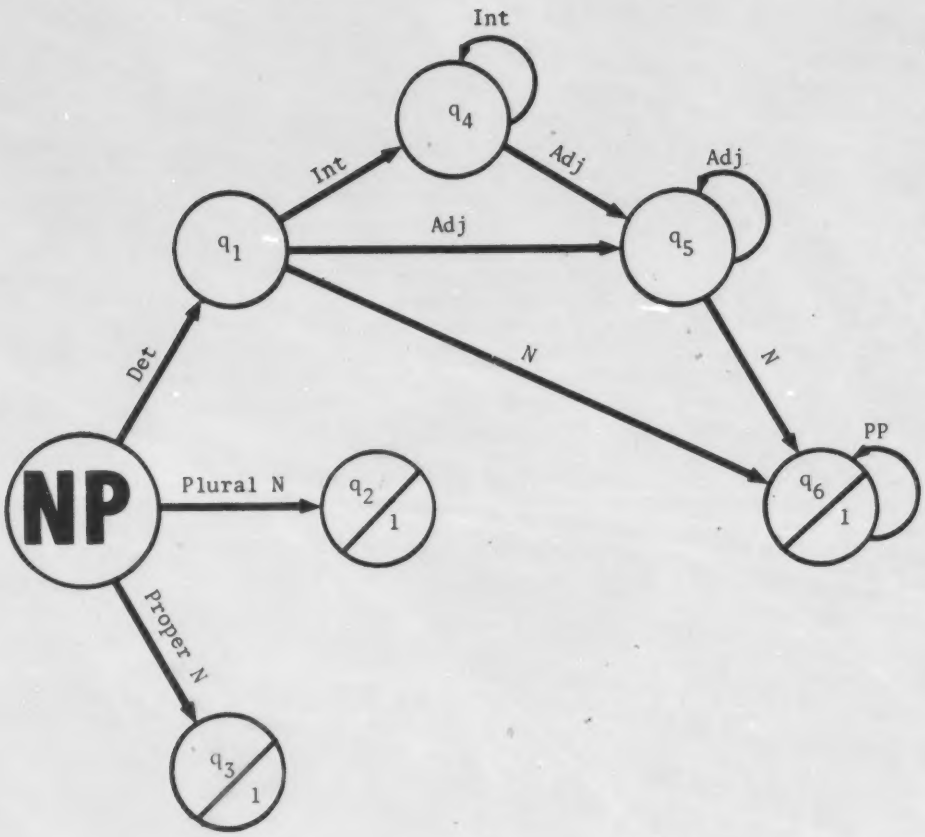


Figure 2. A Sub-Routine for Some Noun Phrases in English

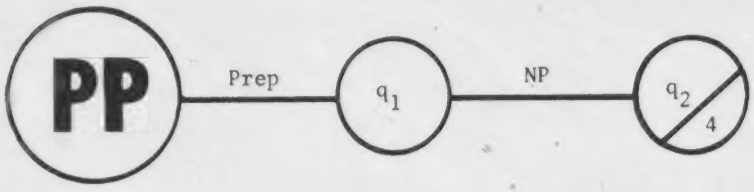
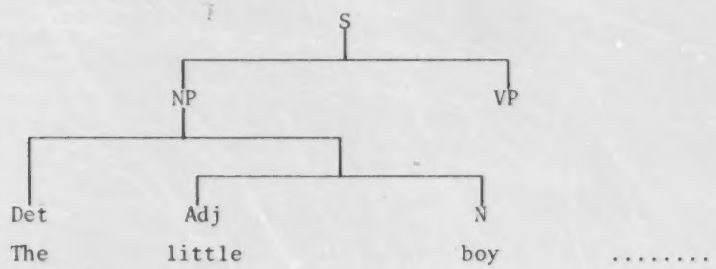


Figure 3. A Sub-Routine for Some Prepositional Phrases in English.

is segmented from the phonetic sequences of [lɪr̆l̩b], [lɪr̆l̩bɔ], [lɪr̆l̩bɔɪ], etc. This is accomplished by virtue of the fact that the word little is the first phonetic sequence that not only matches a lexical entry in the grammar, but also is an adjective which allows a transition in the grammar from state q_1 to q_5. The grammar now anticipates either another

adjective or the head noun to follow. Since little is semantically marked as a size indicator, the grammar anticipates a head noun that pragmatically maps onto a pre-specified physical object which has size and shape dimensions in some extralinguistic context. The next word scanned is boy. (Henceforth, reference to phonetic segmentation is omitted.) Since boy is a noun, it allows the transition from state q_5 to q_6. Since q_6 is a final state for the noun phrase sub-routine (as indicated by the slash and subscript 1 in Figure 2), the grammar may now return processing to the declarative sentence routine provided that no prepositional phrase follows the presumed subject noun phrase. Since the next word is was, which is lexically classified as a form of a stative verb be, the grammar "pops" out of the NP sub-routine and returns control to the declarative sentence routine, where it starts processing the predicate (see Figure 1).

So far, it has recognized and interpreted the noun phrase the little boy, which it has tentatively parsed and interpreted semantically and pragmatically. The parsing is equivalent to an incomplete tree structure as shown below. In addition, the grammar has pragmatically mapped the subject NP onto some referent which is known to be a member



of the set of boys and the sub-set of little boys. On the basis of the declarative sentence generator in Figure 1, it now anticipates a predicate which will tell something about the little boy. The kinds of predicates that are likely are limited by syntactic, semantic, and pragmatic constraints. Syntactically, on the basis of the transition network of Figure 1, the NP may be followed by a transitive verb by taking the transition from q_1 to q_3, or an intransitive verb by taking the transition from q_1 to q_2, etc. The next word scanned by the grammar is was, which allows a transition from q_1 to q_4. This is a final state, but since the word bucked follows, the grammar does not "pop" (i.e. terminate processing in the routine). As soon as bucked off is processed and recognized as a past participle form of a transitive verb, the grammar has specified the passive construction and that an agent of a certain type is likely to follow. Pragmatically the grammar knows that certain four-legged beasts which are ridden by human beings buck. Hence, an agent is expected in a subsequent by-phrase, and certain properties of that agent are anticipated. The effect of this anticipation will be to facilitate processing if correct, and to hinder it if incorrect. Wanat (1974) reported that a prepositional phrase with by, such as by the barn in the sentence, the boy was bucked off by the barn was more difficult to process than a by-phrase such as by the pony in the sentence, the boy was bucked off by the pony. Not only is the by-phrase expected to refer to an agent, but it is pragmatically constrained to mention an agent of a certain sort. The rest of the analysis of Sentence *a* progresses through the by-phrase

and concludes with the processing of the agent noun phrase much the way the subject noun phrase was processed. When processing concludes, the grammar pops from state q_{14} of Figure 1.

Sentence *b* is analyzed in a similar way right up to the past participle, which in *b* is gone away. The possibilities that exist at that point, namely the ones that are specified by the grammar, are that the sentence will terminate, or that it will be followed by a modifier in the form of a prepositional phrase, or other adverbial or string of adverbials. Whatever follows if the sentence does not terminate is not constrained in the way that the agent in the by-phrase of the passive was constrained for sentence *a*. The modifier that is subsequent to the verb in sentence *b* may be a locative such as to his mother's; it may be a time adverbial like by the next day; a manner adverbial like quickly; or an instrumental adverbial like on a horse; etc.

These facts which are conveniently represented in an expectancy grammar explain the data referred to earlier in this paper concerning the differential processing of active and passive sentences. Although the deliberately oversimplified examples that we worked through to this point have attempted to account for language perception, complications of the basic notions illustrated apply equally to perception and production data. For example, combinations of perception and production, as in reading aloud, taking dictation, or having a conversation, all require an expectancy system of the sort illustrated. For this reason it is particularly well-equipped to serve as a framework within which the problems and data of language testing can be discussed.

Another useful extension of a theory of expectancy grammar is suggested by the research of Watzlawick et al. (1967). Their work stresses the fact that human communicative behavior has at least two aspects: there is a *relationship* (affective) aspect of messages concerning how people see each other as people, and there is a *content* (cognitive) aspect encompassing the coding of factual information in the everyday sense (see the remarks above in the section on historical perspective, pp. 41-43). To oversimplify a bit, factors that pertain to the relationship aspect of communication are basically the area of interest encountenanced by sociolinguists; factors that pertain to the content aspect, on the other hand, are characteristically the domain of interest of logicians, cognitive psychologists, and psycholinguists. Relationship information is normally coded by what we term paralinguistic mechanisms, whereas content information is normally coded in segmental phonemes, words, phrases, sentences, etc.

The research of Ogston and Condon (1971) shows an interesting connection between the paralinguistic mechanisms and content level mechanisms of coding. They demonstrated that "as a normal person speaks, his body 'dances' in precise and ordered cadence with the speech as it is articulated. The body moves in patterns of change which are directly proportional to the articulated pattern of the speech stream (p. 153)." And what is perhaps still more interesting: "A hearer's body was found to 'dance' in precise harmony with the speaker. When the units of change in their behavior are segmented and displayed consecutively, the speaker and hearer look like puppets moved by the same set of strings (p. 158)." In commenting on the research of Ogston and Condon, Lenneberg (1971) observed that it is apparently the case "that the flow of movements that constitute motor behavior consists of 'chunks' each having a peculiar program of nervous integration (p. 175)." Lenneberg goes on to observe that,

> the sequences of behavior in animals and man are, under normal conditions, extremely flexible. As the organism moves from situation to situation, the patterns of sequences are constantly readjusted to fit specific demands; the only common denominator that remains between one motor sequence, for example, one episode of catching prey, and the next is a logical principle or, in other words, a generalized pattern. If the individual associates of change of neuromuscular events had to be stored one by one, it is difficult to see how and when the organism would have time to acquire the unique behavioral changes as they occur on one particular occasion, and how instantaneous transformations, which adjust behavior to the imperatives of the moment could be performed without a new trial and error procedure (p. 177).

All of this suggests that the organism possesses a complicated generative mechanism or hierarchy of programs that determines the behavior appropriate to a given situation. An expectancy grammar seems to be a natural mechanism for explaining these facts. Such a grammar seems not only to underlie language behavior in particular, but human behavior in general. That is to say, the human being has internalized a grammar of expectancy which enables him to generate (out of a rich repertoire of "grammatical" routines and sub-routines) unique models to fit particular situations. I have suggested elsewhere that it is reasonable to assume that such grammatical programs themselves are generated and constantly modified by certain principles of learning (Oller, 1971a, 1972c, 1974b, and forthcoming).

CONNECTIONS WITH LANGUAGE TESTING

Within the context of expectancy grammars as models of underlying competence, *a valid language test can be defined as one that activates the expectancy grammar that the learner has internalized.* The extent to which the learner's grammar is able to synthesize and analyze meaningful sequences of elements in the language is an index of his proficiency or competence in the language. A great deal of data from research in second language testing, in particular, shows that some kinds of tests are better than others at activating the learner's internalized expectancy grammar. The arguments considered in this section originated largely in second language proficiency research. However, it should be borne in mind that the conclusions and generalizations from the data have *much* wider applicability.[3]

Among the tests that appear to provide *valid* information about language proficiency are the traditional dictation and the more recently popularized cloze procedure. In various forms, these two and other integrative tests have been advocated by Carroll (1961), Valette (1964, 1967), Spolsky et al. (1968), Oller (1970a; 1973b; forthcoming and references there), Johansson (1974), Angelis (1974), Upshur (1972), Upshur and Palmer (1974), Gradman (1973), Stubbs and Tucker (1974), and many others. One of the indications of the validity of such tests is their strong intercorrelation with each other. A cloze test is based on a visual input that is read by the examinee, while a dictation is based on an auditory input that is heard by the examinee; nevertheless, they tend to correlate at near the .90 level. This means that roughly 81% of the variance on the tests is common variance. Similar results have been achieved with tests

of reading and speaking skills (Oller and Perkins, forthcoming, and Appendix to Oller, forthcoming).

While the typical interpretation of such data by experts prior to the 1970's was that it only indicated test reliability (and more recently Rand, 1972, and Educational Testing Service, 1970), there is reason to believe that such data can only be interpreted as an indication of test validity. In particular, such results show that the two types of tests are probably tapping a common underlying skill. The notion of an "expectancy grammar" offers a sound theoretical basis for explaining the overlap.

Figure 4 illustrates the facts that have been observed in a wide variety of integrative tests, especially the subclass of pragmatic language tests. The areas of the various circles can be taken as rough representations of the variances on different pragmatic tests involving listening, speaking, reading, and writing. The overlap between the circles may be taken as evidence of a basic expectancy grammar that is tapped by all of the tests. While the areas of non-overlap, i.e. the shaded portions of the figure, may perhaps be attributed either to relatively superficial differences in peripheral processing mechanisms (such as hearing and seeing), or to unreliability, or to both, in the cases of individual subjects where the correlation may be practically nil, an explanation can be provided on the basis of disorders in the peripheral processing mechanisms, unusual experiential background (such as only having experienced the language in written form), and the like. For example, a person may be weak in the ability to read English script even though he understands spoken English very well. Similarly, a learner may have acquired considerable skill in deciphering the written form of the language and be quite inept at understanding its spoken form.

Some of the tests which qualify as belonging to the family of pragmatic tests, in addition to those we have already noted, include the Foreign Service Institute's Oral Interview, many reading tasks, essay writing, and a great many other communication tasks.[4] In general, these tests have remarkable characteristics of stability and sensitivity. To illustrate some of their practical and theoretical virtues, we will review only a couple of samples of data from dictation and cloze tests.

A dictation of the sort that is reasonably termed a pragmatic test is one that is administered at a normal conversational rate over segments that challenge the short-term memory span of the examinees. This kind, contrary to much of what the "experts" said during the 1950's and 1960's, has proved repeatedly to be an excellent device for the measurement of language proficiency (Oller, 1970a; Johansson, 1974; Oller and Streiff, 1975), and it also works well as an elicitation device for data concerning specific deficiencies in the internalized grammar of the second language learner (Angelis, 1974). On a dictation as a global proficiency test, the examinee's score is determined by counting the number of deleted words, extraneous insertions, reversals of order, and phonologically mutilated entries.

Of course, more specific "achievement" tests may be constructed by salting the passage with particular sorts of phonological, morphological, lexical, syntactic, semantic, or pragmatic exemplars of rule applications. Many other uses for this integrative testing technique, and the others mentioned earlier, are not difficult to imagine. An important point to remember in the construction of any such test is that the language it represents should be characteristic of the kinds of situations and styles of speech or writing the examinee is apt to encounter in the "real" use

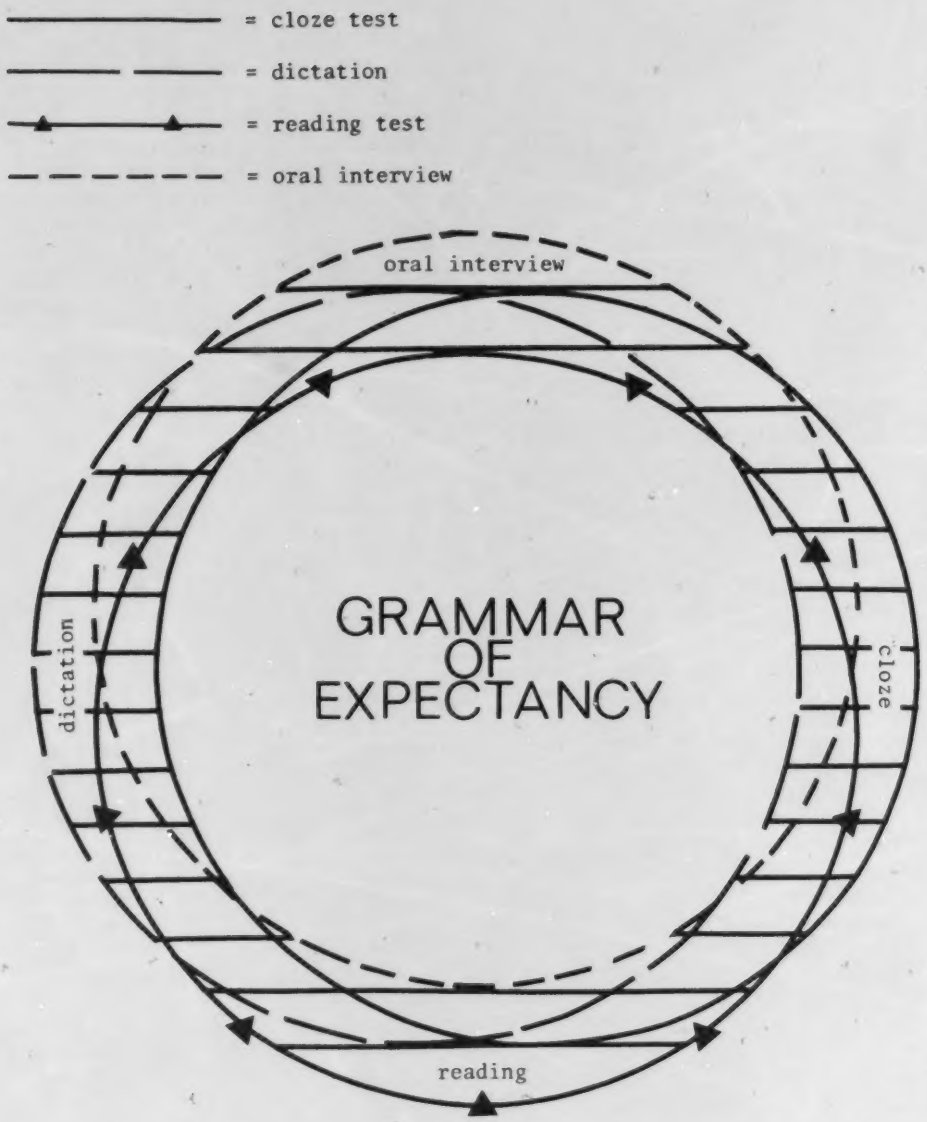


Figure 4. Variance overlap on integrative tests as an indication of an underlying grammar of expectancy.

of the language skills that the test seeks to measure. This is just another way of saying that the activities required by the test, to be valid, must resemble (at a deep level) as accurately as possible the real life activities that they try to predict the examinee's skill in.

A cloze test is constructed by deleting words from a passage of prose. The task set the examinee is to replace the missing items. Probably the simplest and most frequently used method for constructing a cloze test in order to achieve a global estimate of language skill is to delete every sixth or seventh word from a passage of prose of 300 to 350 words in length. The scoring of the test may be done in several ways. The least complicated method is to count the number of words that are restored exactly as they were placed in the original text. Although this method works about as well as any other in determining the readability of a text for native speakers, either one of two other methods will probably work better in the measurement of non-native proficiency. Darnell (1968) recommends a method based on a fairly sophisticated comparison of the non-native's choice on each item against a response frequency analysis for native speakers. The drawback to this scoring technique is its considerable complexity. Having native speakers make judgments of acceptability actually works slightly better than the exact-word scoring method, and is simpler to use than methods requiring response frequency analysis (Darnell, 1968). As in the case of dictation, cloze tests also are adaptable to many different testing purposes. Again, their theoretical claim to validity resides in the fact that they activate the internalized expectancy grammar of the examinee.

Pragmatic tests, such as cloze procedure and dictation, have not been entirely wanting for criticism. It has been argued (Rand, 1972) that they do not provide specific enough information on the precise points of grammar where the examinee may be deficient, whereas discrete-point tests are especially designed to do so. It seems, however, that this criticism is not well-founded. In general, pragmatic tests are much better equipped to provide diagnostic information than discrete-point tests because pragmatic tests elicit errors in contexts where the dynamic aspects of grammar are operative (cf. Angelis, 1974; and Oller, forthcoming).

It has also been argued that pragmatic tests do not clearly differentiate non-native students who can and cannot succeed in college level course work (Rand, 1972). This argument is based on the false assumption that a language test by itself can predict which foreign students will succeed in a college level (or any other) course of study. If language skill were all that were required, no fluent native speaker should ever fail, and this is simply not the case.

Yet another criticism is that integrative tests are "too vague," that one cannot tell precisely what they measure. This, however, is no fault of the tests but is rather characteristic of the skill(s) they seek to measure. Answering the question, "How much language proficiency is necessary to understand what goes on in a college level course" is like answering the question, "How much light is sufficient to find your way out of a forest?"[5] It depends hardly at all on any particular discrete "rays" of light, and the same sort of thing can be said about "points" of language skill. It just happens to be the case, moreover, that pragmatic tests are better suited to the mapping measurement of language proficiency (albeit in a vague way) than discrete-point tests are.

Probably, expectancy grammars are as resistant to mutilations of vari-

ous sorts as is the use of language itself. We can frequently say what we do not mean and still be correctly understood. Or, we may mean what we do not say and still be understood. The well-known and now classic experiments of Miller et al. (1951) for Bell Laboratories are evidence enough of the fact that the speech signal itself may be quite badly distorted and still be understood readily by a listener. The concept of a pragmatically-motivated expectancy grammar suggests a straightforward explanation for this. Similarly, a person who scarcely utters a word that is clearly understandable in isolation may be understood easily when his communications are dealt with in their extralinguistic and linguistic contexts. Thus, pragmatic tests sample the non-native speaker's ability to do what native speakers do in the normal use of language.

SUMMARY

The study of pragmatics as it relates generally to linguistic theory and applied linguistics, and particularly to language testing, has been considered. Although the pragmatic approach may represent a change in the emphasis of current research, it is rooted in a long history of concern for the meaning-oriented aspects of language use and learning. The transformational tradition has largely ignored the pragmatic nature of language until quite recently, but now seems to be moving rapidly in the direction of a pragmatically-motivated theory. The major premise of pragmatics that a whole is greater than the sum of its parts has important consequences for linguistics, applied linguistics, and language testing. Nor is it a new idea. The notion of an "expectancy grammar" is employed as a basis for explaining certain psycholinguistic facts as well as data from language testing research. Pragmatic tests, it is claimed, are superior to the discrete-point type in that they tap the underlying, internalized expectancy grammar of the examinee.

FOOTNOTES

[1]This article incorporates and expands material presented in three earlier papers. The first, "Pragmatic Language Testing: A Theory for Use and a Use for Theory," was an invited lecture presented at Indiana University in January, 1973 at a meeting sponsored by the Committee on Research in Educational Development and Language Instruction. (A version of that lecture appeared in *Language Sciences*, 28, 1973, pp. 7-12.) The second lecture, entitled "Pragmatics," was presented at the University of New Mexico in April, 1973 at a meeting of the Duke City Linguistics Circle. The third lecture, which had the same title as the present paper, was given in San Juan, Puerto Rico in May, 1973 at an International Seminar on Language Testing, jointly sponsored by the AILA Commission on Language Tests and Testing and the Organization of Teachers of English to Speakers of Other Languages (TESOL).

[2]The relevant works of any one of the four men mentioned here would constitute an impressive bibliography. In connection with the notion "pragmatics" and its relation to the philosophy of pragmatism, see Hayden and Alworth (1965) and the selections in it by Pierce, James, and Dewey. For one of the major works on the topic, see Morris (1938). A thorough history of the topic would take us too far off course.

[3]For instance, see Oller and Perkins (1978) for research demonstrating the relevance of expectancy grammar to first language and bilingual con-

texts. (See also Part VI of Oller and Perkins, forthcoming.) Apparently, language proficiency in the sense discussed in this paper is the key factor in a very wide range of educational contexts and especially tests. In fact, one can make a case for the view that intelligence is intrinsically tied to language proficiency in the sense defined, and the tests aimed at the former may really only be measuring the latter (see Oller, 1978).

[4]Pragmatic tests are defined (Oller, forthcoming) as that class of integrative tests meeting two requirements: first, they must require the pragmatic mapping of utterances (or their surrogates) onto extralinguistic context. This can be termed the *meaning* requirement. Second, they must require the processing to take place under temporal constraints. This may be termed the *time* requirement. Integrative tests, on the other hand, are a much broader class of tests defined as the antithesis of discrete-point tests.

[5]I am indebted to the late Dr. Walton Geiger for this metaphor.

References

Aarsleff, H. *The Study of Language in England, 1780-1860.* Princeton, N.J.: Princeton University Press, 1967

Agard, F.B. and H.B. Dunkel. *An Investigation of Second Language Teaching.* Boston, Mass.: Ginn and Co., 1948

Agheyisi, R. and J.A. Fishman. "Language Attitude Studies: A Brief Survey of Methodological Approaches." *Anthropological Linguistics* 12:137-157, 1970

Allen, J.P. and H.G. Widdowson. "Grammar and Language Teaching." In J.P. Allen and S.P. Corder, eds., *Edinburgh Course in Applied Linguistics, Volume 2: Papers in Applied Linguistics.* London: Oxford University Press, 1975, pp. 45-97

Angelis, P. "Listening Comprehension and Error Analysis." In G. Nickel, ed., *AILA Proceedings, Copenhagen 1972, Volume I: Applied Contrastive Linguistics.* Heidelberg: Julius Groos Verlag, 1974, pp. 1-11

Apte, M.L. "'Thank You' and South Asian Languages: A Comparative Sociolinguistic Study." *International Journal of the Sociology of Language* 3:67-89, 1974

Bar-Hillel, Y., ed. *Pragmatics of Natural Languages.* Dordrecht, The Netherlands: D. Reidel Publishing Co., 1971

Bender, M.L., R.L. Cooper, and C.A. Ferguson. "Language in Ethiopia: Implications of a Survey for Sociolinguistic Theory and Method." *Language in Society* 1:215-233, 1972

Blatchford, C.H. "A Theoretical Contribution to ESL Diagnostic Test Construction." *TESOL Quarterly* 5:209-215, 1971

Blickenstaff, C.B. "Musical Talents and Foreign Language Learning Ability." *Modern Language Journal* 47:359-363, 1963

Bloomfield, L. *Language.* New York: Holt, Rinehart and Winston, 1933

Blount, B.G. "Aspects of Luo Socialization." *Language in Society* 1: 235-248, 1972

Boyle, T.A., W.F. Smith, and R.G. Eckert. "Computer Mediated Testing: A Branched Program Achievement Test." *Modern Language Journal* 60: 428-440, 1976

Brière, E.J. "Some Psycholinguistic Considerations Involved in a Language Testing Project." In G.E. Perren and J.L.M. Trim, eds., *Applications of Linguistics: Selected Papers of the Second International Conference of Applied Linguistics, Cambridge, 1969.* Cambridge: Cambridge University Press, 1972, pp. 141-153

Brown, R.W. and M. Ford. "Address in American English." *Journal of Abnormal and Social Psychology* 62:375-385, 1961

Bursill-Hall, G.L. *Speculative Grammars of the Middle Ages: The Doctrine of Prates Orationis of the Modistae.* (Approaches to Semiotics Series, 11.) The Hague: Mouton, 1971

Carranza, M.A. and E.B. Ryan. "Evaluative Reactions of Bilingual Anglo and Mexican American Adolescents toward Speakers of English and Spanish." *International Journal of the Sociology of Language* 6:83-104, 1975

Carroll, J.B. "Knowledge of English Roots and Affixes as Related to Vocabulary and Latin Study." *Journal of Educational Research* 34: 102-111, 1940

_____. "An Appraisal of Language Tests from the Standpoint of the Psychology of Language." *Yearbook of the National Council on Measurements Used in Education, 1951-2*, 9:75-80, 1952

_____. *The Study of Language: A Survey of Linguistics and Other Related Disciplines in America*. Cambridge, Mass.: Harvard University Press & London: Oxford University Press, 1953

_____. "Fundamental Considerations in Testing for English Langugage Proficiency of Foreign Students." In Center for Applied Linguistics, *Testing the English Proficiency of Foreign Students*. Washington, D.C.: Center for Applied Linguistics, 1961, pp. 30-40. Reprinted in H.B. Allen and R.N. Campbell, eds., *Teaching English as a Second Language: A Book of Readings*. New York: McGraw-Hill, 1972

_____. "The Prediction of Success in Intensive Foreign Language Training." In R. Glaser, ed., *Training Research and Education*. Pittsburgh: University of Pittsburgh Press, 1962, pp. 87-136

_____. "A Model of School Learning." *Teachers College Record* 64: 723-733, 1963

_____. "Fundamental Considerations in Testing for English Language Proficiency of Foreign Students." In H.B. Allen, ed., *Teaching English as a Second Language: A Book of Readings*. New York: McGraw-Hill, 1965, pp. 364-372

_____. "Research in Teaching Foreign Languages." In N.L. Gage, ed., *Handbook of Research on Teaching*. Chicago: Rand McNally, 1965

_____. "Psychological Considerations in Setting Aims in Foreign Language Teaching." In V. Steblik and V. Fried, eds., *Rapport: Seminaire internationale sur la differenciation structurale entre les buts de l'enseignement des langues*. Prague: Fédération Internationale des Professeurs de Langues Vivantes (FIPLV), 1968a

_____. "The Psychology of Language Testing." In A. Davies, ed., *Language Testing Symposium: A Psycholinguistic Approach*. (Language and Language Learning Series, 21.) London: Oxford University Press, 1968b, pp. 46-69

_____. "What Does the Pennsylvania Foreign Language Research Project Tell Us?" *Foreign Language Annals* 3:214-236, 1969

_____. "Current Issues in Psycholinguistics and Second Language Teaching." *TESOL Quarterly* 5:101-114, 1971

_____. "Defining Language Comprehension: Some Speculations." In J.B. Carroll and R.O. Freedle, eds., *Language Comprehension and the Acquisition of Knowledge*. New York: Halsted Press, 1972a

_____. *Lectures on English Language Testing and Teaching*. Tokyo: Taikushan Publishing Co., 1972b

_____. "Language Proficiency Tests Developed for the IEA International Study of Achievement in French as a Foreign Language." In B. Spolsky, ed., *Some Major Tests*. (Papers in Applied Linguistics: Advances in Language Testing Series, 1.) Arlington, Va.: Center for Applied Linguistics, 1978, pp. 1-48

_____, A.S. Carton, and C.P. Wilds. *An Investigation of "Cloze" Items in the Measurement of Achievement in Foreign Languages*. Cambridge, Mass.: Laboratory for Research in Instruction, Harvard University Graduate School of Education, 1959

_____ and S.M. Sapon. *Modern Language Aptitude Test*. New York: The

Psychological Corporation, 1959

Cartier, F.A. "Is Learning English worth the Trouble, Miss Fidditch." *Foreign Language Annals* 5:331-340, 1972

Center for Curriculum Development. *Voix et Images de France: Teacher's Manual for Language Skills Tests.* Philadelphia, Pa.: Center for Curriculum Development, 1971

Cherry, C. *On Human Communication.* 2nd ed. New York: John Wiley & Sons, 1965

Chomsky, N. "Three Models for the Description of Language." *IRE Transactions on Information Theory* It-2:113-124, 1956. Reprinted in R.D. Luce, R. Bush, and E. Galanter, eds., *Handbook of Mathematical Psychology.* New York: John Wiley & Sons, 1965

_____. *Syntactic Structures.* (Janua Linguarum Series Minor, 4.) The Hague: Mouton, 1957

_____. *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press, 1965

_____. *Cartesian Linguistics: A Chapter in the History of Rational Thought.* New York: Harper and Row, 1966

_____. *Language and Mind.* New York: Harcourt Brace, 1968. 2nd ed. 1972

_____ and M. Halle. *Sound Patterns of English.* New York: Harper and Row, 1968

Clark, H.H. "The Prediction of Recall Patterns in Simple Active Sentences." *Journal of Verbal Learning and Verbal Behavior* 5:99-106, 1966

Clark, J.L.D. "MLA-Cooperative Foreign Language Tests." *Journal of Educational Measurement* 2:234-244, 1965

_____. *Foreign Language Testing: Theory and Practice.* Philadelphia, Pa.: Center for Curriculum Development, 1972a

_____. "Measurement Implications of Recent Trends in Foreign Language Teaching." In D.L. Lange and C.J. James, eds., *Foreign Language Education: A Reappraisal.* (ACTFL Review Series, 4.) Skokie, Ill.: National Textbook Co., 1972b, pp. 219-257

_____. "Some Considerations in Preparing Test Questions." *Northeast Conference Reports of the Working Committees: 1975.* Middlebury, Vt.: Northeast Conference on the Teaching of Foreign Languages, 1975a, pp. 99-121

_____. "Theoretical and Technical Considerations in Oral Proficiency Testing." In R.L. Jones and B. Spolsky, eds., *Testing Language Proficiency.* Arlington, Va.: Center for Applied Linguistics, 1975b, pp. 10-28

Coffman, W. "Essay Testing." In R.L. Thorndike, ed., *Educational Measurement.* 2nd ed. Washington, D.C.: American Council on Education, 1971

Conway, M.E. "Design of a Separable Transition-Diagram Compiler." *Communications of the Association for Computing Machinery* 6:396-408, 1964

Cooper, R.L. "An Elaborated Language Testing Model." *Language Learning,* Special Issue No. 3:57-72, 1968

_____. "What Do We Learn When We Learn a Language." *TESOL Quarterly* 4: 303-314, 1970

Corder, S.P. "Idiosyncratic Dialects and Error Analysis." *International Review of Applied Linguistics* 9:147-160, 1971

Cox, R.C. and B.G. Sterrett. "A Model for Increasing the Meaning of Standardized Test Scores." *Journal of Educational Measurement* 7:227-228, 1970

Darnell, D.K. *The Development of an English Language Proficiency Test of Foreign Students Using a Clozentropy Procedure.* Boulder, Co.: Department of Speech and Drama, University of Colorado, 1968. [Also available from ERIC Document Reproduction Service--ED 024 039.]

Dewey, J. "Nature, Communication, and Meaning." 1926 ed. reprinted in Hayden and Alworth, eds., 1965, pp. 265-296

Diamond, J. and W. Evans. "The Correction for Guessing." *Review of Educational Research* 43:181-191, 1973

Di Pietro, R.J. *Language Structures in Contrast.* Rowley, Mass.: Newbury House, 1971

Donaldson, M. and R.J. Wales. "On the Acquisition of Some Relational Terms." In J.R. Hayes, ed., *Cognition and the Development of Language.* (Carnegie-Mellon Cognition Series.) New York: John Wiley & Sons, 1970, pp. 235-268

Educational Testing Service. *Handbook: MLA-Cooperative Foreign Language Tests.* Princeton, N.J.: Educational Testing Service, 1965

_____. *Test of English as a Foreign Language: Interpretive Information.* Princeton, N.J.: Educational Testing Service, 1970

_____. *Peace Corps Language Proficiency Tests, Spanish: Spoken Grammar.* Princeton, N.J.: Educational Testing Service, 1971

Einstein, A. "The Common Language of Science." 1951 ed. reprinted in Hayden and Alworth, eds., 1965, pp. 34-37

El-Dash, L. and G.R. Tucker. "Subjective Reactions to Various Speech Styles in Egypt." *International Journal of the Sociology of Language* 6:33-54, 1975

Ferguson, C.A. "Baby Talk in Six Languages." *American Anthropologist* 66 (6, Part 2):103-114, 1964

Fillmore, C. "The Language Teacher as Linguist." Paper presented at the Convention of Teachers of English to Speakers of Other Languages, San Juan, Puerto Rico, May 1973

Firth, J.R. "A Synopsis of Linguistic Theory, 1930-55." *Studies in Linguistic Analysis*, Special Publication of the Philological Society, 1-33, 1957

Fischer, J.L. "Social Influences on the Choice of a Linguistic Variant." *Word* 14:47-56, 1958

Fishman, J.A., ed. *Advances in the Sociology of Language, Volume 1.* (Contributions to the Sociology of Language Series, 1.) The Hague: Mouton, 1971

_____, R.L. Cooper, R. Ma et al. *Bilingualism in the Barrio.* (Language Science Monographs, 7.) Bloomington: Research Center for the Language Sciences, Indiana University, 1971

Gagné, R.M. *The Conditions of Learning.* New York: Holt, Rinehart and Winston, 1965

Gaies, S.J., H.L. Gradman, and B. Spolsky. "Toward the Measurement of Functional Proficiency: Contextualization of the Noise Test." *TESOL Quarterly* 11:51-57, 1977

Gardner, R.C. and W.E. Lambert. *Attitudes and Motivation in Second Language Learning.* Rowley, Mass.: Newbury House, 1972

George, H.V. "Testing--Another Point of View." *English Language Teaching* 16:72-78, 1962

Gradman, H. "Fundamental Considerations in Language Testing." Paper presented at the International Seminar on Language Testing, jointly sponsored by TESOL and the AILA Commission on Language Tests and Testing, held in San Juan, Puerto Rico, May 11, 1973

Greene, J. *Psycholinguistics: Chomsky and Psychology.* (Penguin Science of Behavior Series.) Harmondsworth, England: Penguin Books, Ltd., 1972

Griffin, P., T. Dietrich, and C. Freeman. *Assessing Comprehension in a School Setting.* P. Griffin, ed. (Papers in Applied Linguistics: Linguistics and Reading Series, 3.) Arlington, Va.: Center for Applied Linguistics, 1978

Halliday, M.A.K. "Categories of the Theory of Grammar." *Word* 17:241-292, 1961

_____. *Learning How to Mean: Explorations in the Development of Language.* Amsterdam, The Netherlands: North-Holland Publishing Co., 1977

Hamp, E.P. "What a Contrastive Grammar Is Not, If It Is." In J.E. Alatis, ed., *Georgetown University Round Table 1968.* Washington, D.C.: Georgetown University Press, 1968, pp. 137-147

Hanzeli, V.E. "The Effectiveness of Cloze Tests in Measuring the Competence of Students of French in an Academic Setting." *French Review* 50:865-874, 1977

Harris, D.P. *Teaching English as a Second Language.* New York: McGraw-Hill, 1969

Harris, Z. *Methods in Structural Linguistics.* Chicago: University of Chicago Press, 1951

Hayden, D.E. and E.P. Alworth, eds. *Classics in Semantics.* (Essay Index Reprint Series.) New York: Philosophical Library, 1965

Hayes, A.S., W.E. Lambert, and G.R. Tucker. "Evaluation of Foreign Language Teaching." *Foreign Language Annals* 1:22-44, 1967

Henmon, V.A.C., J.E. Bohan, and C.C. Brigham. *Prognosis Tests in the Modern Foreign Languages.* New York: Macmillan, 1929

Hinofotis, F.B. *An Investigation of the Concurrent Validity of Cloze Testing as a Measure of Overall Proficiency in English as a Second Language.* Doctoral diss. Carbondale, Ill.: Southern Illinois University, 1976

Hjelmslev, L. "La stratification du langage." *Word* 10:163-188, 1954

Holtzman, P.W. "English Language Proficiency and the Individual." In *Selected Conference Papers of the Association of Teachers of English as a Second Language.* Los Altos, Calif.: Language Research Associates' Press, 1967, pp. 76-84

Horton, D.L. and J.J. Jenkins, eds. *The Perception of Language.* New York: Charles E. Merrill Publishing Co., 1971

Houston, S.H. *A Survey of Psycholinguistics.* (Janua Linguarum Series Minor, 98.) The Hague: Mouton, 1971

Hymes, D. "Competence and Performance in Linguistic Theory." In R. Huxley and E. Ingram, eds., *Language Acquisition: Models and Methods.* London: Academic Press, 1971, pp. 3-24

_____. "On Communicative Competence." In J.B. Pride and J. Holmes, eds., *Sociolinguistics.* Harmondsworth, England: Penguin Books, Ltd., 1972

Ingram, E. "Attainment and Diagnostic Testing." In A. Davies, ed., *Language Testing Symposium: A Psycholinguistic Approach.* (Language and Language Learning Series, 21.) London: Oxford University Press, 1968

_____. *Manual for the English Language Battery.* Mimeo. Edinburgh: University of Edinburgh, 1970, pp. 1-20

_____. "A Further Note on the Relationship between Psychological and

Linguistic Theories." *International Review of Applied Linguistics* 9: 335-346, 1971

Ingram, E. "Psychology and Language Learning." In J.P. Allen and S.P. Corder, eds., *Edinburgh Course in Applied Linguistics, Volume 2: Papers in Applied Linguistics.* London: Oxford University Press, 1975, pp. 218-290

Jakobovits, L.A. "A Functional Approach to the Assessment of Language Skills." *Journal of English as a Second Language* 4:63-76, 1969

_____. *Foreign Language Learning: A Psycholinguistic Analysis of the Issues.* Rowley, Mass.: Newbury House, 1970

Jakobson, R. "Boas' View of Grammatical Meaning." *American Anthropologist* 61, Memoirs 89:139-145, 1959

Jesperson, O. *How to Teach a Foreign Language.* London: Allen and Unwin, Ltd., 1904

Johansson, S. "Controlled Distortion as a Language Testing Tool." In J. Qvistgaard, H. Schwarz, and H. Spang-Hanssen, eds., *AILA Proceedings, Copenhagen 1972, Volume III: Applied Linguistics, Problems and Solutions.* Heidelberg: Julius Groos Verlag, 1974, pp. 397-411

Johnson, N. "Linguistic Models and Functional Units of Language Behavior." In S. Rosenberg, ed., *Directions in Psycholinguistics.* New York: Macmillan, 1965, pp. 29-65

Jones, R.L. "The Oral Interview of the Foreign Service Institute." In B. Spolsky, ed., *Some Major Tests.* (Papers in Applied Linguistics: Advances in Language Testing Series, 1.) Arlington, Va.: Center for Applied Linguistics, 1978, pp. 104-115

Jonz, J. "Improving on the Basic Egg: The M-C Cloze." *Language Learning* 26:255-265, 1976

Katz, J.J. and J.A. Fodor. "The Structure of a Semantic Theory." *Language* 39:170-210, 1963. Reprinted in J.A. Fodor and J.J. Katz, eds., *The Structure of Language: Readings in the Philosophy of Language.* Englewood Cliffs, N.J.: Prentice-Hall, 1964

_____ and P.M. Postal. *An Integrated Theory of Linguistic Descriptions.* Cambridge, Mass.: MIT Press, 1964

Kavanagh, J.F. and I.G. Mattingly, eds., *Language by Ear and by Eye: The Relationships between Speech and Reading.* Cambridge, Mass.: MIT Press, 1972

Kirk, S.A., J.J. McCarthy, and W. Kirk. *The Illinois Test of Psycholinguistic Abilities.* Rev. ed. Urbana: University of Illinois Press, 1968

Kolers, P.A. "Eye Voice Span of Response Bias." In Horton and Jenkins, eds., 1971, pp. 17-22

Kuhlwein, Wolfgang, ed., *Linguistics in Great Britain.* 2 vols. Tübingen: Max Niemeyer Verlag, 1971

Kuhn, T.S. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press, 1962

Labov, W. *The Social Stratification of English in New York City.* (Urban Language Series, 1.) Washington, D.C.: Center for Applied Linguistics, 1966

_____. *The Study of Nonstandard English.* Champaign, Ill.: National Council of Teachers of English, 1969

_____, P. Cohen, C. Robins, and J. Lewis. *A Study of the Non-Standard English of Negro and Puerto Rican Speakers in New York City.* New York: Columbia University, 1968

Lado, R. *Linguistics across Cultures: Applied Linguistics for Language*

Teachers. Ann Arbor: University of Michigan Press, 1957

_____. "English Language Testing: Problems of Validity and Administration." *English Language Teaching* 14:153-161, 1960

_____. *Language Testing: The Construction and Use of Foreign Language Tests*. London: Longman, Green, and Co., 1961 & New York: McGraw-Hill, 1964

Lamb, S. *Outline of Stratificational Grammar*. Berkeley: University of California Press, 1966

_____. "Outline of Stage 6: Parts I, II, III." Papers presented at the Conference on Stratificational Linguistics, jointly sponsored by the Summer Institute of Linguistics and the University of Washington, held in Seattle, Washington, August 1973

Lambert, W.E. "A Social Psychology of Bilingualism." *Journal of Social Issues* 23,2:91-109, 1967

_____, R.C. Gardner, R. Olton, and K. Tunstall. "A Study of the Roles of Attitudes and Motivation in Second-Language Learning." In J.A. Fishman, ed., *Readings in the Sociology of Language*. The Hague: Mouton, 1968, pp. 473-491

Lamendella, J. "On the Irrelevance of Transformational Grammar to Second Language Pedagogy." *Language Learning* 19:255-270, 1969

Lashley, K.S. "The Problem of Serial Order in Behavior." In L.A. Jeffress, ed., *Cerebral Mechanisms in Behavior*. New York: John Wiley & Sons, 1951, pp. 112-136. Reprinted in S. Saporta and J. Bastian, eds., *Psycholinguistics: A Book of Readings*. New York: Holt, Rinehart and Winston, 1961, pp. 180-197

Lenneberg, E.A. "The Importance of Temporal Factors in Behavior." In Horton and Jenkins, eds., 1971, pp. 174-184

Levenston, E.A. "Aspects of Testing the Oral Proficiency of Adult Immigrants to Canada." In L. Palmer and B. Spolsky, eds., *Papers on Language Testing 1967-74*. Washington, D.C.: TESOL, 1975, pp. 66-76.

Levin, H. and E.L. Kaplan. "Listening, Reading, and Grammatical Structure." In Horton and Jenkins, eds., 1971, pp. 1-16

Liberman, A. "Some Results of Research on Speech Perception." *Journal of the Acoustical Society of America* 29:117-23, 1957

Lieberman, P. *Intonation, Perception, and Language*. (Press Research Monographs, 38.) Cambridge, Mass.: MIT Press, 1967

Malinowski, B. *Coral Gardens and Their Magic*. London: Allen and Unwin, Ltd., 1935. Reprinted by Indiana University Press (Bloomington), 1965

Marso, R.N. "The Influence of Test Difficulty upon Study Efforts and Achievement." *American Educational Research Journal* 6:621-632, 1969

McCawley, J.D. "The Role of Semantics in Grammar." In E. Bach and R. Harms, *Universals in Linguistic Theory*. New York: Holt, Rinehart and Winston, 1968

Mehler, J. "Some Effects of Grammatical Transformations on the Recall of English Sentences." *Journal of Verbal Learning and Verbal Behavior* 2:346-351, 1963

Miller, G.A. "Some Psychological Studies of Grammar." *American Psychologist* 17:748-762, 1962

_____, H. Heise, and J.C.R. Licklider. "The Intelligibility of Interrupted Speech." *Journal of the Acoustical Society of America* 22: 167-173, 1950

Morgan, W.J. "A Clinical Approach to Foreign Language Achievement." In A.A. Hill, ed., *Georgetown University Round Table 1953*. Washington, D.C.: Georgetown University Press, 1953, pp. 15-21

Morris, C. *Foundations of the Theory of Signs.* (International Encyclopedia of Unified Science 1, No. 2.) Chicago: University of Chicago Press, 1938

Moskowitz, G. *The Foreign Language Teacher Interacts.* New York: Association for Productive Teaching, 1970

National Association for Foreign Student Affairs. *English Proficiency Chart.* Washington, D.C.: National Association for Foreign Student Affairs, 1971

Neisser, U. *Cognitive Psychology.* New York: Appleton-Century-Crofts, 1967

Newcomb, W.B. "A Directed Graph Model for English Syntax." Unpublished ms. Rochester, N.Y.: General Dynamics, 1963

Newcomer, P. et al. "The Construct Validity of the Illinois Test of Psycholinguistic Abilities." *Journal of Learning Disabilities* 8: 220-231, 1975

Ogston, S. and E. Condon. "Speech and Body Motion Synchrony." In Horton and Jenkins, eds., 1971

Oller, D.K. and R.E. Eilers. "Phonetic Expectation and Transcription Validity." *Phonetica* 31:288-304, 1975

Oller, J.W., Jr. "Dictation as a Device for Testing Foreign-Language Proficiency." *UCLA Workpapers in TESL* 4:37-41, 1970a. Reprinted in *English Language Teaching* 25:254-259, 1971

_____. "Transformational Theory and Pragmatics." *Modern Language Journal* 54:504-507, 1970b

_____. "Language Communication and Second Language Learning." *UCLA Workpapers in TESL* 4:99-107, 1970c

_____. *Coding Information in Natural Languages.* The Hague: Mouton, 1971a

_____. "Contrastive Analysis, Difficulty, and Predictability." (Paper presented at the Pacific Conference on Contrastive Linguistics and Language Universals, Honolulu, Hawaii.) *Working Papers in Linguistics* 3, No. 4, 1971b. Reprinted in *Foreign Language Annals* 6:95-106, 1972

_____. "Assessing Competence in ESL: Reading." *TESOL Quarterly* 6: 313-323, 1972a

_____. "Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in English as a Second Language." *Modern Language Journal* 56:151-158, 1972b

_____. "Transfer and Interference as Special Cases of Induction and Substitution." *International Journal of Psycholinguistics* 1:24-33, 1972c

_____. "Discrete Point Tests versus Tests of Integrative Skills." In Oller and Richards, eds., 1973, pp. 184-198, 1973a

_____. "Cloze Tests of Second Language Proficiency and What They Measure." *Language Learning* 23:105-118, 1973b

_____. "Expectancy for Successive Elements: Key Ingredient to Language Use." *Foreign Language Annals* 7:443-452, 1974a

_____. "On the Generation and Modification of Grammars." (Paper presented at the Summer Meeting of the Linguistic Society of America, Amherst, Massachusetts, July 1974.) In A. Makkai and V.B. Makkai, eds., *The First LACUS Forum 1974.* Columbia, S.C.: Hornbeam Press, 1974b, pp. 393-402

_____. "Pragmatic Mappings." (Paper presented at the Conference on Stratificational Linguistics, jointly sponsored by the Summer Institute of Linguistics and the University of Washington, held in Seattle, Washington, August 1973.) *Lingua* 35:333-344, 1975a

_____. *Research with Cloze Procedure in Measuring the Proficiency of Non-Native Speakers of English: An Annotated Bibliography.* Albuquerque, N.M.: University of New Mexico, 1975b. [Also available from ERIC Document Reproduction Service--ED 104 154.]

_____. "Evidence for a General Language Proficiency Factor: An Expectancy Grammar." *Die Neueren Sprachen* 76:165-174, 1976

_____. "The Language Factor in the Evaluation of Bilingual Education." (Paper presented at the Georgetown University Round Table on Languages and Linguistics: International Dimensions of Bilingual Education, Washington, D.C., March 1978.) In J.E. Alatis, ed., *Georgetown University Round Table 1978.* Washington, D.C.: Georgetown University Press, 1978

_____. *Language Tests at School: A Pragmatic Approach.* London: Longman, forthcoming

_____ and N. Inal. "A Cloze Test of English Prepositions." *TESOL Quarterly* 5:315-326, 1971

_____ and K. Perkins, eds. *Language in Education: Testing the Tests.* Rowley, Mass.: Newbury House, 1978

_____ and K. Perkins, eds. *Research in Language Testing.* Rowley, Mass.: Newbury House, forthcoming

_____ and J.C. Richards, eds. *Focus on the Learner: Pragmatic Perspectives for the Language Teacher.* Rowley, Mass.: Newbury House, 1973

_____ and B.D. Sales. "Conceptual Restrictions on English: A Psycholinguistic Study." *Lingua* 23:209-232, 1969

_____, B.D. Sales, and R.V. Harrington. "A Basic Circularity in Traditional and Current Linguistic Theory." *Lingua* 22:317-328, 1969

_____ and A.V. Streiff. "Dictation: A Test of Grammar Based Expectancies." *English Language Teaching* 30:25-36, 1975. Also in R.L. Jones and B. Spolsky, eds., *Testing Language Proficiency.* Arlington, Va.: Center for Applied Linguistics, 1975, pp. 71-88

Olson, D.R. "Language Use for Communicating, Instructing, and Thinking." In J.B. Carroll and R.O. Freedle, eds., *Language Comprehension and the Acquisition of Knowledge.* New York: Halsted Press, 1972

O'Neill, D. "The Creation of Language by Children." In J. Lyon and R.J. Wales, eds., *Psycholinguistic Papers.* Edinburgh, Scotland: University Press, 1966, pp. 99-115

Osgood, C.E. "Motivational Dynamics of Language Behavior." In *Nebraska Symposium on Motivation.* Lincoln, Neb.: University of Nebraska Press, 1957a

_____. "A Behavioristic Analysis of Perception and Language as Cognitive Phenomena." In I. de Sola Pool, ed., *Contemporary Approaches to Cognition: A Report of a Symposium at the University of Colorado, 1955.* Cambridge, Mass.: MIT Press, 1957b, pp. 75-118

_____. *Speculation on the Structure of Interpersonal Intentions.* Mimeo. Urbana: Department of Psychology, University of Illinois, 1966

Pack, E.C. "The Effects of Testing upon Attitude towards the Method and Content of Instruction." *Journal of Educational Measurement* 9: 141-144, 1972

Packard, R.G. "Models of Individualized Instruction: The Search for a Measure." *Educational Technology,* 11-14, August 1972

Parent, P.P. and F.P. Veidt. "Program Evaluation: Accountability." In D.L. Lange, ed., *Britannica Review of Foreign Language Education, Volume 3.* Chicago, Ill.: Encyclopaedia Britannica, 1971, pp. 311-339

Palmer, Adrian S. "Testing Communication." *International Review of Applied Linguistics* 10:35-45, 1972

Paulston, C.B. "Linguistic and Communicative Competence." *TESOL Quarterly* 8:347-362, 1974

Perren, G.E. "Testing Ability in English as a Second Language: 3. Spoken Language." *English Language Teaching* 22:22-29, 1967

_____. "Testing Spoken Language: Some Unsolved Problems." In A. Davies, ed., *Language Testing Symposium: A Psycholinguistic Approach.* (Language and Language Learning Series, 21.) London: Oxford University Press, 1968

_____. "Specifying the Objectives: Is a Linguistic Definition Possible?" *English Language Teaching* 25:132-139, 1971

Pike, E.V. "A Test for Predicting Phonetic Ability." *Language Learning* 9:35-41, 1959

Pike, L.W. *An Evaluation of Present and Alternative Item Formats for Use in the Test of English as a Foreign Language.* Princeton, N.J.: Educational Testing Service, 1973

Pilliner, A.E.G. "Subjective and Objective Testing." In A. Davies, ed., *Language Testing Symposium: A Psycholinguistic Approach.* (Language and Language Learning Series, 21.) London: Oxford University Press, 1968, pp. 19-35

Pimsleur, P. *Manual: Language Aptitude Battery.* New York: Harcourt Brace, 1966

_____. *Pimsleur Language Aptitude Battery.* New York: Harcourt Brace, 1966

_____, L. Mosberg, and A.L. Morrison. "Student Factors in Foreign Language Learning." *Modern Language Journal* 46:160-170, 1962

Poulter, V.L. "Computer-Assisted Laboratory Testing." *Modern Language Journal* 53:561-564, 1969

Rand, Earl J. "Integrative and Discrete-Point Testing at UCLA." *UCLA Workpapers in TESL* 6:67-78, 1972

Reichenbach, H. *Elements of Symbolic Logic.* New York: Macmillan, 1947

Reichling, Anton. "Principles and Methods of Syntax: Cryptanalytic Formalism." *Lingua* 10:1-17, 1961

Rice, F.A. "The Foreign Service Institute Tests Language Proficiency." *The Linguistic Reporter* 1:4, May 1959

Richards, J.C. "A Non-Contrastive Approach to Error Analysis." (Paper presented at the Convention of Teachers of English to Speakers of Other Languages, San Francisco, March 1970.) *English Language Teaching* 25:204-219, 1971a. Also in Oller and Richards, eds., 1973, pp. 96-113

_____. "Error Analysis and Second Language Strategies." *Language Sciences* 17:12-22, 1971b. Also in Oller and Richards, eds., 1973, pp. 136-144

Roberts, K. "The Interaction of Normative Associations and Grammatical Factors in Sentence Retention." (Paper presented at the Midwestern Psychological Association Meeting, Chicago, 1966.) [As cited in Levin and Kaplan (see Horton and Jenkins, eds., 1971).]

Robins, R.H. "General Linguistics in Great Britain, 1930-1960." In C. Mohrmann, F. Norman, and A. Sommerfelt, eds., *Trends in Modern Linguistics.* Utrecht, The Netherlands: Spectrum, 1963, pp. 11-37. Reprinted in Kuhlwein, Vol. II, 1971, pp. 3-33

Rosenabum, Y., E. Nadel, R.L. Cooper, and J.A. Fishman. "English on Keren Kayemet Street." In J.A. Fishman, R.L. Cooper, and A.W. Conrad,

eds., *The Spread of English: The Sociology of English as an Additional Language*. Rowley, Mass.: Newbury House, 1977, pp. 179-194

Rulon, P.J. "On the Validity of Educational Tests." *Harvard Educational Review* 16:290-296, 1946

Russell, B. *An Inquiry into Meaning and Truth*. New York: W.W. Norton and Co., 1940

Rutherford, W.E. *Modern English*. New York: Harcourt Brace, 1968

Salomon, E. "A Generation of Prognosis Testing." *Modern Language Journal*. 38:299-303, 1954

de Saussure, F. *Course in General Linguistics*. W. Baskin, trans. C. Bally and A. Sechehaye, eds. New York: Philosophical Library, 1912

Shoemaker, D.M. "Toward a Framework for Achievement Testing." *Review of Educational Research* 45:127-147, 1975

Sinclair, H. "Sensorimotor Action Patterns as a Condition for the Acquisition of Syntax." In R. Huxley and E. Ingram, eds., *Language Acquisition: Models and Methods*. London: Academic Press, 1971, pp. 121-130

Sinclair-de-Zwart, H. and J. Flavell. "Developmental Psycholinguistics." In D. Elkind and J.H. Flavell, eds., *Studies in Cognitive Development: Essays in Honor of Jean Piaget*. New York: Oxford University Press, 1969, pp. 315-336

Skinner, B.F. *Verbal Behavior*. New York: Appleton-Century-Crofts, 1957

Smith, P.D. *A Comparison of the Cognitive and Audiolingual Approaches to Foreign Language Instruction*. Philadelphia, Pa.: Center for Curriculum Development, 1970

Spolsky, B. "Language Testing--The Problem of Validation." *TESOL Quarterly* 2:88-94, 1968

_____. "Reduced Redundancy as a Language Testing Tool." In G.E. Perren and J.L.M. Trim, eds., *Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics, Cambridge, 1969*. Cambridge: Cambridge University Press, 1972, pp. 383-390

_____. "What Does It Mean to Know a Language, or How Do You Get Someone to Perform His Competence?" In Oller and Richards, eds., 1973, pp. 164-176

_____, B. Sigurd, M. Sato, E. Walker, and C. Arterburn. "Preliminary Studies in the Development of Techniques for Testing Overall Second Language Proficiency." *Language Learning*, Special Issue No. 3, 79-101, 1968

_____, P. Murphy, W. Holm, and A. Ferrel. "Three Functional Tests of Oral Language Proficiency." *TESOL Quarterly* 6:221-235, 1972

Steiner, F. "Behavioral Objectives and Evaluation." In D.L. Lange, ed., *Britannica Review of Foreign Language Education, Volume 2*. Chicago, Ill.: Encyclopaedia Britannica, 1970, pp. 35-78

Stevens, K.N. "Toward a Model of Speech Perception." *Journal of the Acoustical Society of America* 32:47-55, 1960

Stubbs, J.B. and G.R. Tucker. "The Cloze Test as a Measure of English Language Proficiency." *Modern Language Journal* 58:239-241, 1974

Taylor, W.L. "Cloze Procedure: A New Tool for Measuring Readability." *Journalism Quarterly* 30:414-438, 1953

Tursi, J.A., ed. *Foreign Languages and the "New" Student*. (Reports of the Working Committees of the Northeast Conference on the Teaching of Foreign Languages.) New York: Northeast Conference on the Teaching of Foreign Languages, 1970

Uhlenbeck, E.M. "Some Further Remarks on Transformational Grammar." *Lingua* 17:263-316, 1967

Upshur, J.A. "Language Proficiency Testing and the Contrastive Analysis Dilemma." *Language Learning* 12:123-128, 1962

_____. "Productive Communication Testing: A Progress Report." In G.E. Perren and J.L.M. Trim, eds., *Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics, Cambridge, 1969*. Cambridge: Cambridge University Press, 1972, pp. 435-441

_____ and A. Palmer. "Measures of Accuracy, Communicativity, and Social Judgments for Two Classes of Foreign Language Speakers." In A. Verdoodt, ed., *AILA Proceedings, Copenhagen 1972, Volume II: Applied Sociolinguistics*. Heidelberg: Julius Groos Verlag, 1974

Vachek, J. *The Linguistic School of Prague*. Bloomington: Indiana University Press, 1966

Valette, R.M. "The Use of the Dictée in the French Language Classroom." *Modern Language Journal* 48:431-434, 1964

_____. *Modern Language Testing: A Handbook*. New York: Harcourt Brace, 1967. Rev. ed. 1977

_____. "Some Conclusions to Be Drawn from the Pennsylvania Study." *National Association of Language Laboratory Directors Newsletter* 3, iii:17-19, 1969

_____ and R.S. Disick. *Modern Language Performance Objectives and Individualization: A Handbook*. New York: Harcourt Brace Jovanovich, 1972

Vernon, P.E. *Secondary School Selection*. London: Methuen, 1960

Wanat, S. "Effects of Language Organization on Perceptual Processing." In A. Verdoodt, ed., *AILA Proceedings, Copenhagen 1972, Volume II: Applied Sociolinguistics*. Heidelberg: Julius Groos Verlag, 1974

_____ and H. Levin. "The Eye-Voice Span: Reading Efficiency and Syntactic Predictability." In H. Levin et al., eds., *The Analysis of Reading Skill: A Program of Basic and Applied Research*. Ithaca, N.Y.: Cornell University, 1968, pp. 237-253

Watzlawick, P., J. Beavin, and K. Jackson. *Pragmatics of Human Communication*. New York: W.W. Norton and Co., 1967

Wilds, C.P. "The Oral Interview Test." In R.L. Jones and B. Spolsky, eds., *Testing Language Proficiency*. Arlington, Va.: Center for Applied Linguistics, 1975, pp. 29-44

Woods, W.A. "Transition Network Grammars for Natural Language Analysis." *Communications of the Association for Computing Machinery* 13:591-602, 1970

_____. *An Experimental Parsing System for Transition Network Grammars*. (Report No. 2362.) Cambridge, Mass.: Bolt, Beranek, and Newman, Inc., 1972

Zirkel, P.A. et al. "The Validation of Parallel Testing." *Psychology in the Schools* 2:153-157, 1974